IDENTIFICATION OF SYSTEMS OF UNKNOWN ORDER

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
MARTIN W. SCHWARTZ
1972

This is to certify that the

thesis entitled

IDENTIFICATION OF SYSTEMS

OF UNKNOWN ORDER

presented by

Martin W. Schwartz

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in SYS

Gerald L Park

Major professor

Date April 6, 1972

ABSTRACT

IDENTIFICATION OF SYSTEMS OF UNKNOWN ORDER

By

Martin W. Schwartz

The primary purpose of this work is to determine the order of a discrete-time system from input and output observations.

Several tests, based on the properties of dynamic systems, are given. The best of them is the determinants of the models' controllability matrices normalized by the product of their input coefficients. This and the other new tests compare favorably with existing tests, based on the properties of the least squares estimator, on simulated examples and observed data taken from an electric power system.

Theoretical properties of the least squares estimator when the order of the model is not equal to the minimal order of the system are also developed. When the model order is higher than the system order in a noise free system, extra poles and zeros are added to the transfer function; all of these are on the same side of the unit circle. When noise is present, the estimator still converges, but to an unknown realization. It is also shown that the small sample estimates might theoretically differ greatly from the noise-free estimates in such situations. However, numerical results are quite close.

Martin W. Schwartz

Variants of the standard least squares estimator are developed and compared. A starting algorithm for on-line estimation, based on matrix pseudo-inverses is also developed.

IDENTIFICATION OF SYSTEMS OF UNKNOWN ORDER

By

Martin W. Schwartz

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering
and Systems Science

1972

To Joyce

who makes it all worthwhile

ACKNOWLEDGEMENTS

The author wishes to express his sincere gratitude to Dr. G.L. Park, his thesis advisor for his continued assistance and encouragement during the course of this research.

Thanks are also due to Dr. R.O. Barr, Dr. J.S. Frame, Dr. A. Bacopolous, and Dr. J.B. Kreer for their interest and critical reviews of the thesis and their excellent teaching. The author also appreciates the financial support awarded through the Consumers Power Company-Michigan State University Cooperative Research Agreement.

TABLE OF CONTENTS

		Page
	ABSTRACT	
	LIST OF FIGURES	v
	LIST OF TABLES	vi
Chapter		
I	INTRODUCTION	1
	1.1 Motivation	1
	1.2 Background	6
	1.2.1 Equivalent Systems and Identifiability	6
	1.2.2 Identification of Linear Systems	14
II	LEAST SQUARES ALGORITHMS	18
	2.1 Statement of the Problem	18
	2.2 Large-sample Properties	21
	2.2.1 Uncorrelated Noise	21
	2.2.2 Correlated Noise	23
	2.3 Small-sample Properties	24
	2.4 Systems of Unknown Order	26
	2.4.1 No Noise	27
	2.4.2 Noise	30
	2.4.3 Tests for the Order of the System	33
	2.5 Normalized Least Squares	39
	2.6 Summary	41
III	NUMERICAL COMPUTATIONS	42
	3.1 Description of the Algorithms Used	42
	3.2 Description of the Examples	46
	3.3 Computed Results	48
IV	CONCLUSIONS	68
	4.1 Thesis Results	68
	4.2 Future Development	69
	APPENDIX	70
	BIBLIOGRAPHY	77

LIST OF FIGURES

Figure		Page
1	Norm of θ vs. Extra Pole	32
2	Determinants of Controllability Matrices	51
3	Ratio of Controllability Determinants (full matrix lse)	52
4	Order Tests	53
5	Mean Square Residuals	54
6	Mean Squared Residual for Example 2	60
7	Noramlized Controllability Determinant for Example 2	61
8	Mean Squared Residual -- Example 3	65
9	Controllability Determinants -- Example 3	66

LIST OF TABLES

Table		Page
1	Equivalent Realizations of a System	13
2	Order Tests for Example 1	50
3	Critical Values for the F-test	55
4	F-test for Example 1	56
5	Estimates of the System Order	57
6	Order Test Summary for Example 2	57
7	F-tests for Example 2	58
8	Order Tests for Example 2	59
9	Poles and Zeros for Example 2	63
10	Order Test Summary for Example 3	63
11	Order Tests for Example 3	64
12	F-test for Example 3	67
13	Poles and Zeros for Example 3	67

CHAPTER I

INTRODUCTION

1.1 Motivation

The following problem arose in a study of an electric power system: find a simple equation which represents the behavior of a high-voltage tieline connecting two areas. Such a mathematical model would be incorporated in simulation models used in design of the system and would be studied to possibly provide better control of the system. The standard theoretical treatments of these tielines (E2) requires knowledge of parameters which are often not available and some simplifying assumptions which were questioned. Thus, it was decided to measure the power and frequency (the relevant variables) on a tieline and form a model using these measurements.

This is an example of a common problem we will call the identification problem; namely, to obtain information about a complicated real-world dynamic system from discrete-time or sampled measurements of the relevant input and output variables. Using known information and enough simplifying assumptions to make the problem computationally feasible, a mathematical model with unknown parameters is constructed, and the parameters are estimated. If the model is accurate, some properties of the system can be determined within certain accuracies. Specifically, consider a

single-input, single-output (SISO) dynamic system which is linear and time-invariant. If the order n is known, the problem is to estimate the coefficients of the difference equation:

$$(1.1.1) \qquad y_t = \Sigma_1^n a_i y_{t-i} + \Sigma_1^n b_i u_{t-i}$$

where y is the output, u the input, and $y_t = y(t)$. For example, a model for one tieline could be

$$(1.1.2) \qquad f_t = .59f_{t-1} + .405f_{t-2} + .003p_{t-1} + .002p_{t-2}$$

where f is frequency and p is power. Sometimes the input is a test signal, sometimes not; the system may be initially at rest, or may not; and we may have knowledge or can make assumptions about the measurement errors and plant noise. In each situation some method such as least squares, Gauss-Markov (L1), maximum likelihood (A4), Kalman-Ho (K2), or stochastic approximation (M3) is available to generate the estimates.

Going back to the specific problem of modeling the load-frequency behavior of an electric power tieline, the measurements must be taken during normal operation, and noise is present as measurement error in both input (power) and output (frequency) and as an inherent stochastic element of the system (the system loads and configuration are partly random) (P1). While the model can be taken as linear and time-invariant during a short period under normal ranges of the variables, several factors arise that violate the usual assumptions of existing methods. First, the stochastic assumptions necessary to insure unbiased estimates are violated. Second, we do not have sufficient information about the system

to obtain the order of the model or relations between the parameters. Third, we have very limited information about the characteristics of the noise. Finally, the output and, to a lesser extent, the input are almost constant.

The fourth factor is important. The usual methods for testing the fit of a model calculate the residuals; if they are small, the model is considered good. But here the residuals alone do not constitute a sufficient test for the validity of the model because any reasonable identification scheme will yield small residuals. For example, if a constant model equal to the nominal value of 60.000Hz is taken, one set of data has an rms residual of .02% of the estimate. But our goal is to derive some information about the structure of the system, and not just to estimate the output. Thus, the almost constant data implies that we must look for additional criteria to use in testing the model fit.

The constancy causes additional problems. Many identification methods involve solution of a set of equations of the form $Y = XA$ derived from (1.1.1). Several columns of X will contain output values (which in the above example vary from 59.980 to 60.017Hz), and these columns will be almost linearly dependent. Therefore, the solution will be very sensitive to small perturbations in the data and to round-off errors in the numerical technique. Another aspect of the problem is that the high intercorrelation of the columns of X makes it very difficult to determine empirically a reasonable order for the model. This is important because, as shown in Section 1.2.1, different models may exhibit identical input-output behavior and one might be stable and the other unstable,

one controllable and the other uncontrollable, or have other divergent properties.

The following example from numerical analysis (I2) shows this problem in a slightly different context. When solving the differential equation

$$dy/dt = -y, \quad y(0) = y_o$$

with a third-order difference approximation (of a specified form) for the derivative, the difference equation

$$y_t + (3/2 + 3h)y_{t-1} - 3y_{t-2} + \tfrac{1}{2}y_{t-3} = 0$$

is obtained, where h is the sampling interval. Three distinct solutions of the form $y_t = a^t$ are obtained. All of the a's satisfy the equation

$$a^3 + (3/2 + 3h)a^2 - 3a + \tfrac{1}{2} = 0 \ .$$

For sufficiently small h, one eigenvalue is less than -2. Thus the difference solution becomes unbounded while the true solution approaches zero. The problem is due solely to the addition of the extraneous eigenvalues caused by too large a model order.

Our primary motivation is solving the practical problem of how to choose a model. In doing so we will not always be rigorous from a probabilistic interpretation. This is deliberate and is common in developing numerical techniques; Powell (P2), when discussing methods for finding the maximum of a function $F(x_1,x_2,\ldots,x_n)$, says:

> The lack of definiteness in stating the conditions on $F(x_1,x_2,\ldots,x_n)$ is deliberate, because we are describing the current state of optimization, and it happens that the current state is not a logical structure of theorems. Instead it has developed from an assortment of numerical methods which have been devised because real problems had to be solved, and at present the actual success of the algorithms is far ahead of any theoretical predictions.

The model-order problem is much less developed than the maximization problem; numerical, rather than theoretical, results for the model order problem are emphasized.

There are some new theoretical results in this thesis: Theorem 2 on the asymptotic convergence of least-squares estimators with correlated noise; Theorem 3 on small-sample properties with white noise, and Theorem 4 on the location of the additional poles with too-high a model order and no noise. The algorithm in Chapter III for the pseudo-inverse of the sum of matrices of a certain form is also new, and a variant of the least squares algorithm which we call the reduced least squares algorithm has not been explicitly discussed in the literature. However, for the solution of the practical problem, the main contribution is the application of the theory of equivalent representations to the study of the identification problem.

The nature of the problem precludes a logical structure of theorems at the present time. The rest of this chapter and most of the next develop theory necessary to the understanding of the tests for model order and their numerical comparison. Relevant literature is reviewed when the appropriate subject is discussed and is often juxtaposed with the new material; this is true, in particular, for the section on the model-order problem.

1.2 Background

In this section we will define and discuss several concepts necessary to define the problem more clearly and to compare each method with others. The first part distinguishes the different ways we can represent a system and discusses related questions; although the discussion will be for linear, discrete, time-invariant SISO systems, most of the ideas are valid for multivariate and time-varying ones. The next part describes the different loss or error criteria we can use for identification. The last part is a brief outline of several significant identification algorithms and how they become more elegant and efficient as we have more information and more control over how the information is obtained.

It is necessary in what follows to distinguish between a real-world system, a mathematical system, and a mathematical model of a system. A real-world system is, of course, what we find in nature and almost always lacks desirable properties (such as true time-invariance) that mathematical systems can have. In Section 1.2.1 we will distinguish between a mathematical system and a model or representation of it. This distinction is more subtle and is more important because, although we are developing algorithms with the intent that they will be applied to real-world problems, much of the time we will have to talk about models of mathematical systems.

1.2.1 Equivalent realizations and identifiability

In the remaining chapters we talk about "the order of the system" and "the order of the model" as if they were well-defined

concepts, and it is important that they be well defined. The most obvious definition (for either concept) would be the largest lag in the difference equation, i.e. the parameter n in equation (1.1.1), but this is not sufficient because it does not distinguish between difference equations which have identical input-output behavior.

This problem can be thought of (G1, K1) as pole-zero cancellation of the transfer function T(z) relating input to output. The discrete transform function is the z-transform (O1) of the differential equation representation of the system. The z-transform is the discrete version of the LaPlace transform and has many of the same properties. An example of the use of the z-transform and the corresponding derivation using the difference equation is in the Appendix. If we derive T(z) from theoretical considerations and get a pole-zero pair, then it is well known (G1) that if this pole is unstable, then the real system will act as though it was unstable even though classical z- (or LaPlace) transform theory tells us that we may cancel that term in the mathematical model and look at just the other poles. The reasons for this deviant behavior are that real components are not linear over all possible input-output values (C2) and that they do not exhibit exact pole-zero cancellation (K1).

A simple example (C2), using a continuous time dynamic model, demonstrates this. Let us set up the system with total transfer function $T(s) = 1/(s-1)$ using two components in series such that the first has $T_1(s) = 1/(s+1)$ and the second $T_2(s) = (s+1)/(s-1)$, so that $T(s) = T_1(s)T_2(s)$. With zero initial

conditions and unit step input, theoretically $y(t) = 1 - \exp(-t)$ the same as if we had used only one integrator. Real components, however, are linear for only a limited range of their input and output. Thus, after a while, the output of the first component or the input of the second would be overloaded, and the observed behavior would deviate markedly from the theoretical. This is true even if all the components are perfect; if the pole and zero do not match exactly, or if there are slight perturbations in the system, the effect is increased.

However, we are deriving $T(z)$ from observed data, not theoretical considerations. As explained later in the section the input-output data gives us information only about the controllable, observable part of the system, and this part has no pole-zero pairs. If everything is exact, and we take the model order to be larger than the system order, the model transfer function will have pole-zero pairs (see Section 2.4.1) and cancelling these will yield the transfer function of the (controllable, observable part of the) system; here we cancel the pole-zero pairs because any occurence of such pairs from input-output data is due to a flaw in the identification algorithm or an error we have made because we cannot identify those pairs which do cancel. If there are stochastic elements, then we do not get exact pole-zero pairs because of the identification errors, but we would like to guarantee that the behavior is not much different than the behavior of the noise-free system.

We now have two reasons for examining concepts related to equivalent representations. First, we want to have an unambiguous

statement of what "the order of the system" means. Second, we must be able to compare two systems and/or models and be able to say when their behavior is identical or close.

There are three different ways of representing the relation between the input and output which we shall use: difference equation, transfer function, and state-space model. Each of these has advantages and disadvantages, some of which are discussed below. But first let us explicitly state them. The difference equation representation is

$$y_t = \Sigma_1^n a_i y_{t-i} + \Sigma_1^n b_i u_{t-i} \tag{1.2.1}$$

where the a_i and b_i are constants and n is the maximum lag. Its z-transform is (01)

$$T(z) = \frac{b_1 z^{n-1} + b_2 z^{n-2} + \ldots + b_n}{z^n - a_1 z^{n-1} - \ldots - a_n} \quad . \tag{1.2.2}$$

There are several standard state-space representations of the difference equation (1.2.1) (01), one of which is as follows: let

$$\begin{aligned} h_1 &= b_1 \\ h_2 &= b_2 + a_1 h_1 \\ &\vdots \\ h_n &= b_n + a_1 h_{n-1} + \ldots + a_{n-1} h_1 \quad . \end{aligned} \tag{1.2.3}$$

Now define the state variables:

$$\begin{aligned} x_1(t) &= y_t \\ x_2(t) &= x_1(t+1) - h_1 u_t \\ &\vdots \\ x_n(t) &= x_{n-1}(t+1) - h_{n-1} u_t \end{aligned} \tag{1.2.4}$$

Then the state-space model consists of state- and output-equations

$$X(t+1) = AX(t) + Bu_t$$

(1.2.5)

$$y_t = CX(t)$$

where

$$X(t) = \mathrm{col}(x_1(t),\ldots,x_n(t))$$

(1.2.6)

$$A = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & 0 & 1 \\ a_n & a_{n-1} & \cdots & a_2 & a_1 \end{bmatrix}$$

$$B = \mathrm{col}(h_1,\ldots,h_n)$$

$$C = (1,0,\ldots,0)$$

Two important concepts derived from state-space representations that we will need for the model-order problem are controllability and observability (G1, O1, K1). Formally, a state-space realization is controllable (observable) if its controllability matrix Q_c (observability matrix Q_o)

$$Q_c = (B,\ AB,\ldots,A^{n-1}B)$$

(1.2.7)

$$Q_o = (C^T, A^TC^T,\ldots,(A^T)^{n-1}C^T)$$

has full rank. Intuitively, it is observable if we can see what is happening internally (in the state vector) by looking at the output. It is controllable if we can manipulate the state using the input. The difference equation representation contains information only about those parts of the system which are both

controllable and observable. In the above example, the differential equation does not show the output of the first integrator or the input to the second, and it does not allow us to manipulate them so that the desired input-output relation holds for all values.

The state-space representation (1.2.3)-(1.2.6) has observability matrix equal to the identity. This can readily be seen because C^T is the first column of the identity and $(A^T)^k C^T$ is the first column of $(A^T)^k$ which is the kth column of A^T and is equal to the kth column of the identity. To illustrate several ideas in this section we will use Example 1, Section 3.2, whose difference equation form is

$$(1.2.8) \quad y_t = .8y_{t-1} + .39y_{t-2} - .27y_{t-3} - .5u_{t-1} + .5u_{t-2} + .1u_{t-3}$$

Its z-transform is

$$(1.2.9) \quad T(z) = \frac{-.5z^2 + .5z + .1}{z^3 - .8z^2 - .39z + .27}$$

$$= \frac{-.5(z-1.17082)(z + .17082)}{(z-.9)(z+.6)(z-.5)}$$

And for the state-space representation we are using

$$(1.2.10) \quad A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -.27 & .39 & .8 \end{bmatrix} \qquad B = \begin{bmatrix} -.5 \\ .1 \\ -.015 \end{bmatrix} \qquad C = (1\ 0\ 0)$$

We are now ready to show that (1.2.8) and the other representations are not unique, i.e. there are other difference equations which have identical input-output behavior. The material is partially contained in several references (A3, B5, R2, Z1, Z2) which differ among themselves in definitions; ours will also differ,

but will generally follow the later sources. We shall define a system as an input sequence and its corresponding output sequence. The triplet (A, B, C), defined above, is called a (state-) realization of the system. Any system obviously has many realizations such as (A, B, C) and its similar realizations $(T^{-1}AT, T^{-1}B, CT)$ where T is any nonsingular matrix of appropriate size. We shall say that two realizations, S and S', are equivalent if, given an initial state x_o of S, there is an initial state x_o' of S' such that

$$y(t; x_o, u) = y'(t; x_o', u).$$

More simply, two realizations are equivalent if the same input string applied to both yields the same output string. Obviously, two similar realizations are equivalent, not all equivalent realizations are similar; while all similar realizations have the same dimension state vector, not all equivalent realizations do. If there is no realization with a smaller-dimension state vector, then the realization is called minimal. All minimal realizations are similar. Finally, the order of the system is the dimension of the minimal realization. We will also call the coefficient vector of the difference equation form (1.2.1) a realization, keeping in mind that formally we mean the one obtained from (1.2.1) by using (1.2.3)-(1.2.6).

If the order of the system is m, then there is no difference equation with fewer than m lags which realizes the system. If we do not know m and try models with other maximum lags, n, then the models will also realize the system. Thus we

will use the following terminology to differentiate between the maximum lag and the order of the system: by "the order of the model" we will mean the largest lag of the difference equation in the realization being used. Table 1 illustrates the difference between "the order of the system" and "the order of the model" by showing several equivalent realizations of Example 1, where the order of the system is 3.

Table 1. Equivalent Realizations of a System

order of the model	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
3	.8	.39	-.27	---	-.5	.5	.1	---
4	.8	.39	-.27	0	-.5	.5	.1	0
4	1.8	-.41	-.66	.27	-.5	1.0	-.4	-.1
4	-.466	-.657	.140	.090	-.5	.332	.267	.033

The following facts about minimal and non-minimal realizations will be useful:

1) There is no pole-zero cancellation in the transfer function form (1.2.2) of the minimal realization; non-minimal realizations have pole-zero cancellation. (1.2.2) shows that the model of order 3 has no cancellation; the models of order 4 in Table 1 have additional factors of z, $(z-1)$, and $(z - .334021)$, respectively. This is the main advantage of form (1.2.2), since factoring the numerator and denominator of $T(z)$ is the easiest way to determine if a realization is minimal. The main disadvantage is that $T(z)$ does not account for the initial conditions unless they are zero.

2) The difference equation form (1.2.1) of a minimal realization is unique; this is not true of non-minimal realizations of fixed order as Table 1 illustrates.

3) Given any realization of a system and its controllability and observability matrices Q_c and Q_o, then the order of the system is $\text{rank}(Q_o^T Q_c)$. Realization (1.2.3)-(1.2.6) has the advantage that the order of the system it represents is equal to the rank of Q_c, making calculations easier.

We will finish the discussion of equivalent realizations in Section 2.4, after some other relevant material has been presented.

The last question in this section concerns identifiability (B2, S2). A model is identifiable if its coefficients can be determined uniquely from the input-output data. This requirement of uniqueness is desirable to inhibit the estimator from oscillating between two correct coefficient vectors (i.e. between two realizations). If m is less than n, there are many possible solutions and the uniqueness property is not satisfied. But there can be convergence without uniqueness (B1); for example, the algorithm can converge to the realization with minimum norm. This is the case here (see Section 2.4.2), and we do not care very much which realization is approached. Thus we will procede to identify the parameters even if they are not identifiable in the formal sense. They will be identifiable in practice.

1.2.2 Identification of linear systems

If we were extremely fortunate, we would be able to use a test signal for the input and observe the output without any noise.

For such situations, the Ho-Kalman algorithm (K2) provides an elegant method for determining the order of the system and the parameter vector. Let the input be a pulse and let

$$H = \begin{bmatrix} y_1 & y_2 & \cdots \\ y_2 & y_3 & \cdots \\ . & . & \cdots \\ . & . & \cdots \\ . & . & \cdots \end{bmatrix} \tag{1.2.11}$$

H is a Hankel matrix, i.e. its (i,j)th entry depends only on $(i+j)$. Let its principal leading submatrix of dimension n be denoted by H_n. If the minimal order of the system is m, then

$$\det(H_n) = 0 \quad \text{if and only if} \quad n > m. \tag{1.2.12}$$

This gives a method for determining the order of the system and a realization of it:

$$\begin{aligned} A &= H_m^{-1}(sH_m), \text{ where } s \text{ is the shift operator} \\ B &= \text{col}(1, 0, \ldots, 0) \\ C &= (y_1, y_2, \ldots, y_m) \end{aligned} \tag{1.2.13}$$

The algorithm generalizes easily to the situation where there are r inputs and p outputs. Then H is a block Hankel matrix with $r \times p$ blocks, and the above expressions for A, B, and C are pre- and post-multiplied by "editing" matrices, composed of the identity and zeros, which are easy to compute. If there is no noise, but a pulse input cannot be used, the method does not work because it is impossible to set up a matrix which has the properties of H which are needed. But there is still no problem because we can set up a system of linear equations of appropriate size and

solve the parameters. On the other hand, there are problems which have not been completely solved, whenever noise is present, no matter what the input.

If there is noise, in general the equations are not consistent, so it is necessary to choose some criterion in making the estimates. The most common criterion is least squares, where the sum of the squared differences between the model estimates and the observed values is to be minimized. This method, elaborated in Chapter II, is the one we have chosen because it has several advantages. It is easy to compute the estimates, and it is essentially distribution free, i.e. it is not necessary to know the distribution of the stochastic elements a priori; it is unbiased as long as the noise is uncorrelated. Other methods, such as maximum-likelihood and the Bayes' method, which maximize probability density functions, require known density functions of the residuals.

A set of variants of the least squares method is weighted least squares. Regular least squares solves a set of equations of the form $Y = XA$ derived from (1.1.1). Then weighted least squares solves $WY = WXA$, where W is a positive definite matrix. A well-known example of a weighting matrix is the inverse of the covariance matrix of the residuals. This variant is called the Gauss-Markov or Best Linear Unbiased (BLUE) estimator. Another variant is the instrumental variables method (A3), and a third is discussed in Section 2.5.

These are the most common criteria in identification algorithms. There are some other criteria which cannot be used with a single estimate, but which can be used to compare different

estimates. For example, we might want the parameter estimates to be near their true values. Since we do not know the true values, parameter errors are not a basis for an identification algorithm, but we can compare the parameter estimates from different estimators. For dynamic systems, one way to compare them (which is independent of model order) is to compare the steady state response to a unit-step input. As shown in the Appendix,

$$y_{ss} = (\Sigma\, b_i)/(1 - \Sigma\, a_i) \tag{1.2.14}$$

where the a_i and b_i are the parameters in the difference equation (1.1.1). We can also compare the estimates of the poles and zeros of the transfer function of a dynamic system. This topic is discussed in detail in Section 2.4.3.

CHAPTER II

LEAST SQUARES ALGORITHMS

In this chapter we will discuss least-squares estimation of discrete, single-input, single-output systems using different assumptions about the noise, when the order of the system is known and when it is not. Two variations of the standard technique are given: a new one which has improved small sample properties, and a weighted one. We will also discuss tests for the order of the system.

2.1 Statement of the Problem

A deterministic system of order n can be represented in discrete time as the difference equation

$$x_t = \Sigma_1^n a_i x_{t-1} + \Sigma_1^n b_i v_{t-i} \tag{2.1.1}$$

where input v and output x are given at equally spaced times. If stochastic elements are present, we shall write the equation to include modeling errors as

$$y_t = \Sigma_1^n a_i y_{t-i} + \Sigma_1^n b_i u_{t-i} + e_t \tag{2.1.2}$$

where e_t will be explicitly defined as appropriate for each theorem. After N observations, (2.1.2) is usually written in matrix form as

(2.1.3) $$Y_N = \phi_N \theta + E_N, \quad \text{where}$$

$Y_N = \text{col}(y_{n+1}, y_{n+2}, \ldots, y_{n+N})$, an $N \times 1$ output vector

$E_N = \text{col}(e_{n+1}, e_{n+2}, \ldots, e_{n+N})$, an $N \times 1$ noise vector

$\theta = \text{col}(a_1, \ldots, a_n, b_1, \ldots, b_n)$, the $2n \times 1$ parameter vector,

and

$$\phi_N = \begin{bmatrix} y_n & \cdots & y_1 & u_n & \cdots & u_1 \\ y_{n+1} & \cdots & y_2 & u_{n+1} & \cdots & u_2 \\ \vdots & & \vdots & \vdots & & \vdots \\ y_{n+N-1} & \cdots & y_N & u_{n+N-1} & \cdots & u_N \end{bmatrix}$$

an $N \times 2n$ observation matrix. The least squares estimator (lse) θ_N of θ is given by

(2.1.4) $$\theta_N = \phi_N^+ Y_N$$

where $+$ denotes the Moore-Penrose generalized inverse. If $N \geq 2n$ and ϕ_N has full (column) rank, then,

(2.1.5) $$\theta_N = (\phi_N^T \phi_N)^{-1} \phi_N^T Y_N$$

The proof is straight-forward using Lagrange multipliers; the errors E_N drop out of the solution because their derivatives with respect to the parameters are zero.

The large and small sample properties of the lse are given in the next two sections; there it is shown that it is biased for a finite number of samples even when it is asymptotically unbiased. The variation we call the reduced lse, which is unbiased for finite samples, can be obtained by deleting entries in Y_N, E_N, and ϕ_N as follows: let

$$Y_K = \text{col}(y_{n+1}, y_{2(n+1)}, \ldots, y_{K(n+1)})$$

and make the corresponding changes in E_N and ϕ_N so that

$$Y_K = \phi_K \theta + E_K \tag{2.1.6}$$

When this is done, any y_t, $t = 1,\ldots,N = K(n+1)$, will appear just once in (2.1.6) while it usually appears $(n+1)$ times in (2.1.3). For example, let $n = 2$ and $N = 7$. Then

$$Y_N = \begin{bmatrix} y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} y_2 & y_1 & u_2 & u_1 \\ y_3 & y_2 & u_3 & u_2 \\ y_4 & y_3 & u_4 & u_3 \\ y_5 & y_4 & u_5 & u_4 \\ y_6 & y_5 & u_6 & u_5 \\ y_7 & y_6 & u_7 & u_6 \\ y_8 & y_7 & u_8 & u_7 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \end{bmatrix}$$

for the full matrix lse, and

$$Y_K = \begin{bmatrix} y_3 \\ y_6 \\ y_9 \end{bmatrix} = \begin{bmatrix} y_2 & y_1 & u_2 & u_1 \\ y_5 & y_4 & u_5 & u_4 \\ y_8 & y_7 & u_8 & u_7 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_6 \\ e_9 \end{bmatrix}$$

for the reduced matrix lse. The reduced estimator is

$$\theta_K = \phi_K^{+} Y_K \tag{2.1.7}$$

Section 2.4 discusses the lse's when the true order of the system is not known and studies the practical problem of choosing a model order; Section 2.5 gives another variation -- normalized least squares.

2.2 Large-sample Properties

In this section the asymptotic properties of the lse of system (2.1.2) are derived under uncorrelated and correlated noise. The asymptotic properties of the reduced matrix form are the same as for the full matrix form.

2.2.1 Uncorrelated noise

The simplest stochastic case is independent system noise, i.e. where the random variables e_t are independent and identically distributed (iid) and are independent of the input and output. This means that there are no observation errors, but only an inherent stochastic behavior in the system. Astrom (A2) states and sketches the proof (with minor errors) that in this case the least squares estimator converges in the mean square. The proof is sketched below because some intermediate results are used in the rest of the chapter.

Theorem 1. (Astrom) Let the system be given by (2.1.2) and assume:

1) the residuals e_t are iid with zero mean, have all moments finite, and are independent of the input and output;
2) the system is stable;
3) $\lim \Sigma_1^N \frac{1}{N} u_t$ is finite and $\lim \Sigma_1^N \frac{1}{N} u_t u_{t-s} = R_u(s)$ is finite for $s = 0,1,\ldots$;
4) the matrix whose (i,j)th element is $R_u(i-j)$ is positive definite;
5) the order of the system is known.

Then the least squares estimator (2.1.4) converges to θ in mean square.

The assumptions of this theorem are common in identification: the first characterizes the errors and will be generalized in Theorem 2; Assumption 2 guarantees that the output is bounded (bounded output should be sufficient for the proof); Assumptions 3 and 4 insure that the input is well behaved (for example, essentially constant inputs such as a unit step or impulse will not work for least squares identification); and the last assumption guarantees uniqueness. The finite difference representation is used for identification because it gives the unknown parameters in terms of directly measurable variables. When the (minimal) order is known, this representation is unique.

Outline of proof (for more details see the proof of Theorem 2 in the appendix): Let

$$B_N = \frac{1}{N} \phi_N^T E_N \quad \text{and} \quad C_N = \frac{1}{N} \phi_N^T \phi_N \tag{2.2.1}$$

Assumptions 2, 3, and 4 imply that $\lim B_N = 0$ and $\lim C_N = C$ exists and is positive definite. Define

$$Z_N = C_N \theta_N' \tag{2.2.2}$$

where $\theta_N' = \theta_N - \theta$ is the error of the N-th estimate. Then

$$Z_N = \frac{1}{N} \phi_N^T E_N = B_N \tag{2.2.3}$$

It is now possible to show that

$$E(Z_N) = 0 \quad \text{and} \quad E(Z_N^T Z_N) = \frac{s}{N} C,$$

where $s = \text{var}(e_t)$. Now $E(Z_N) = C \lim \theta_N'$, so the lse is unbiased. Finally,

$$\lim E(\theta_N'^T \theta_N') = \lim E(Z_N^T Z_N) = \lim \frac{s}{N} \operatorname{tr}(C),$$

so the estimate converges in mean square.

2.2.2 Correlated noise

We will now generalize the results to allow e_t to be system noise correlated to observation noise.

Theorem 2. Let the system be given by (2.1.2) and assume:

1) The residuals e_t are identically distributed with zero mean, all moments finite, $E(e_t e_{t+s}) = 0$ for $s > n$, and are independent of the noise-free input and output.

2) - 5) same as in Theorem 1.

Then the lse (2.1.4) converges to $(\theta + C^{-1}B)$ in mean square, where C is defined by (2.2.1) and

$$(2.2.4) \quad B = \operatorname{col}(E(y_t e_{t+1}),\ldots,E(y_t e_{t+n}), E(u_t e_{t+1}),\ldots,E(u_t e_{t+n}))$$

The proof is given in the appendix. It is similar to the proof of Theorem 1, but now

$$E(Z_N) = B \neq 0 \quad \text{and} \quad E(Z_N^T Z_N) = \frac{1}{N} \text{ ("weighted C").}$$

The term $C^{-1}B$ is the asymptotic bias, which is unknown in most practical situations.

We conclude this section with two comments. First, the assumption that $E(e_t e_{t+s}) = 0$ for $s > n$ is necessary. If it did not hold, the actual order of the system would be greater than n, as it would be if the numerator of the transfer function was higher than the degree of the denominator. Second, if there is

independent observation noise, but no plant noise, then (2.2.4) can be replaced by

$$(2.2.5) \qquad B = \mathrm{col}(-a_1 s_y, \ldots, -a_n s_y, -b_1 s_u, \ldots, -b_n s_u)$$

where s_u and s_y are the variances of the input and output noise respectively.

2.3 Small Sample Properties

While it is theoretically important to have desirable properties when an arbitrary amount of data is available, in practical situations it is more important to know the properties when only a limited amount of data is available. But small-sample properties are more difficult to obtain because they often involve non-linear transformations. Thus, while the proofs of Theorems 1 and 2 (for large samples) did not involve the distribution of e_t, the negative results of the following theorem (for small samples using the full matrix lse) are proved explicitly only for an example involving a particular distribution and specific values for N and n.

Theorem 3. Under the assumptions of Theorem 1, the full matrix lse (2.1.5) may be biased for finite samples, while the reduced matrix lse (2.1.7) is unbiased.

Proof: From (2.2.2) and (2.2.3) we see that

$$(2.3.2) \qquad \theta_N' = \phi_N^+ E_N$$

e_t is independent of all u_s and all e_s and y_s, $s \neq t$. So in the reduced matrix formulation, each element of E_K is independent of each element of ϕ_K, which implies that $E(\phi_K^+ E_K) = 0$,

and θ_K is unbiased. On the other hand, for the full matrix formulation, E_N and ϕ_N^+ are generally dependent. Let us take the simplest example, $n=1$ and $N = 2$. After some algebra, we obtain

$$(2.3.3) \qquad \theta_N' = \frac{1}{e_2u_1+u_1(ay_1+bu_1)-y_1u_2} \begin{bmatrix} e_2(u_2-au_1) \\ e_2(ay_1+2bu_1+e_2) \end{bmatrix}$$

Thus,

$$(2.3.4) \qquad E(a') = \int_{-\infty}^{\infty} \frac{c_3 e}{c_1 + c_2 e} f(e)de$$

where c_1, c_2, and c_3, obtained from (2.3.3), are independent of e_2, and $f(e)$ is the density function of the error. The expected value of b' is similar, but has an additional term, quadratic in e, in the numerator. We will not attempt to find the class of density functions which give unbiased results; but if e_t is uniformly distributed on $(-d,d)$, then

$$(2.3.5) \qquad E(a') = \frac{1}{d}\left[\frac{d}{c_1} + \frac{c_2}{2c_1^2} \log \frac{c_2 - c_1 d}{c_2 + c_1 d}\right] \neq 0, \quad \text{almost surely.}$$

This is the only bias we will explicitly find, but if either n or N is increased, (2.3.4) becomes the multiple integral of the product of a joint density function and a rational function in several variables, so the full matrix lse will give biased small sample estimates most of the time.

Of course, there is a price to pay for the unbiased reduced matrix lse: the convergence factor for the full matrix estimator is $\frac{1}{N}$, while it is $\frac{1}{K} = \frac{n+1}{N}$ for the reduced formulation.

For correlated noise, the situation is less clear. First, it is impossible to eliminate either small or large sample bias because E_K and ϕ_K^+ are dependent for both estimators. Second, it is more difficult to compare the standard estimates, because each estimator has a different weighting matrix. However, numerical results in Chapter III often indicate better results using the reduced lse. It is possible that either the small sample bias or the variance of the estimator is reduced.

The only other results on small-sample estimates are in Hurwicz (H2). He shows that the least squares estimates are biased for a first order free response system when a finite number of samples are available and gives specific values of the bias under very particular circumstances.

2.4 Systems of Unknown Order

The discussion so far has assumed that the order of the system is known; if the system is simple it is usually possible to determine the order from physical, economic, or other relevant theoretical considerations. But if the system is complex, it may not be possible to determine the order from theory. One reason is the problem of aggregation -- even when it is possible to model each component of a system, the total may not act like the sum of the components. This phenomenon has been noted in engineering (C1) and economics (T1). Another reason is that the order of a "complete" model (if it is possible to construct one) may be in the thousands, while the dominant behavior is of low order. Thirdly, the order of the observable, controllable part may be less than

the sum of the orders of the subsystems (G1).

Thus we are motivated to study parameter estimation when the order of the model is not that of the system. First, least squares estimation when the order of the model is higher than the order of the system is presented. Then tests for the order are presented. While not completely general, we will assume that the system has the same form as the model, but has a different order; this assumption will allow us to use several tools to solve the model order problem. Let the order of the system be m so that it may be represented as

$$y_t = \Sigma_1^m c_i y_{t-i} + \Sigma_1^m d_i u_{t-i} + e'_t \tag{2.4.1}$$

and let the order of the model be n so that

$$y_t = \Sigma_1^n a_i y_{t-i} + \Sigma_1^n b_i u_{t-i} + e_t \tag{2.4.2}$$

Further, we will assume that n is larger than m unless otherwise specified.

2.4.1 No noise

Using the material from Section 1.2.1, we see that there is not a unique difference equation if $n > m$. If $n-m = p$, then multiplying both the numerator and denominator of the transfer function $T(z)$ by the same arbitrary p-th degree polynomial yields an exact n-th order realization of the system. Two such realizations are

$$\theta = \text{col}(c_1, c_2, \ldots, c_m, 0, \ldots, 0, d_1, \ldots, d_m, 0, \ldots, 0) \ , \tag{2.4.3}$$

where the c_i and d_i are the coefficients in the minimal realization (2.4.1), and the least squares realization described below which is due to Soderstrom (S1). Table 1 shows examples of fourth order realizations of a third order system; the second row in the table is realization (2.4.3), and the last row is the least squares model.

With no noise, the rank of ϕ is $m+n$, so $\phi^T\phi$ is not invertible. The lse becomes

$$\theta_N = \phi_N^+ Y_N = (\phi_N^T \phi_N)^+ \phi_N^T Y_N \tag{2.4.4}$$

The least squares realization has the smallest norm of any n-th order realization. To characterize the lse we need some new notation. First, rearrange the columns of ϕ and the elements of θ so that the input and output at the same time are adjacent. Then

$$\phi_o = \begin{bmatrix} y_n & u_n & y_{n-1} & u_{n-1} & \cdots & y_1 & u_1 \\ \vdots & & & & & & \vdots \\ y_{n+N-1} & u_{n+N-1} & y_{n+N-2} & u_{n+N-2} & \cdots & y_N & u_N \end{bmatrix} \tag{2.4.5}$$

$$\theta_o = \mathrm{col}(x \quad 0_{2p}) \tag{2.4.6}$$

where

$$x = (c_1, d_1, \ldots, c_m, d_m)$$

and θ_k is the null vector with k elements. Now define

$$\theta_i = \text{col}(0_{2i-2} \quad -1, \quad 0 \quad x \quad 0_{2p-2i}), \quad i = 1,\ldots,p$$

$$\theta_o = \text{col}(a_1,b_1,\ldots,a_n,b_n)$$

(2.4.7)

$$H = \text{col}(\theta_1^T,\ldots,\theta_p^T)$$

$$B = \text{col}(g_1,\ldots,g_p)$$

where G is obtained from the factorization of $T(z)$

$$(2.4.8) \qquad T(z) = \frac{b_1 z^n + b_2 z^{n-1} +\ldots+ b_n z}{z^n - a_1 z^{n-1} -\ldots- a_n}$$

$$= \frac{d_1 z^m +\ldots+ d_m z}{z^m -\ldots- c_m} \cdot \frac{z^p + g_1 z^{p-1} +\ldots+ g_p}{z^p + g_1 z^{p-1} +\ldots+ g_p}$$

θ can be written as

$$(2.4.9) \qquad \theta = \theta_o + g_1\theta_1 +\ldots+ g_p\theta_p = \theta_o + H^T G$$

From (2.4.1) it is clear that $H\phi_o^T = 0$, so that premultiplying by H yields the solution

$$(2.4.10) \qquad G = -(HH^T)^{-1}H\theta_o$$

Numerical examples of this computation are given in Section 3.2.

One property of the real system that the model should reflect is stability or instability. The model will be unstable if one of the extra poles of $T(z)$ lies outside the unit circle, even though the actual system is stable. We want to insure that the algorithm used to estimate the coefficients does not introduce instability into the model; this is especially important if there are stochastic elements present because then the best we can expect is that the pole-zero pairs will be approximately, but not exactly,

equal. Obviously, (2.4.3) yields a stable realization because $G = 0_p$. The situation is not as neat for the lse, as the following theorem shows.

Theorem 4. Let a system be given by (2.4.1) and its model by (2.4.2), where $n-m = p > 0$, and assume that e_t and e_t' are identically zero. Then the lse (2.4.4) introduces p additional pole-zero pairs to the transfer function (2.4.8). Either all these pairs are located inside the unit circle or all are located outside the unit circle. In particular, if $p = 1$, the extra pair is stable.

The proof of Theorem 4 is in the appendix.

Our attention has been focused on the case when n is greater than m because there does not seem to be any practical problem when n is smaller. The relevant theoretical work, called the reduction of dynamic systems, (A1, D1, M2) has assumed that the system is known and all that is desired is a lower order approximation to it. Our situation is somewhat different because all we have are input and output values, but it does not appear to be much different. We get a projection from m dimensional space to n dimensional space which does not seem to effect stability or controllability. In fact, it is sometimes difficult to distinguish between the reduced and unreduced models both in our identification problem (see Section 3.3) and in a similar method for the reduction of known systems (A1).

2.4.2 Noise

There is a great gap between the theoretical development and the practical results when there is noise and $n > m$. The

practical results are amazingly good, while the limited theory says that we have no right to expect anything very close. The basic problem is that the noise-free ϕ_N matrix does not have full rank, and the observed, noisy ϕ_N has full rank almost surely (a.s.). Thus, if $C = \lim \frac{1}{N} \phi_N^T \phi_N$ is invertible (this limit of a sequence of non-singular matrices is not necessarily non-singular), then Theorems 1 and 2 are still valid (with minor changes). C does have full rank a.s. if there are only independent observation errors. To show this let

$$\phi_N = D_N + P_N \tag{2.4.11}$$

where D_N is the deterministic part and P_N are the observation errors. Then

$$E(C_N) = E(\frac{1}{N} D_N^T D_N) + E(\frac{1}{N} P_N^T P_N) = D + P \tag{2.4.12}$$

D has rank (n+m), but P is a diagonal matrix of rank 2n,

$$P = \text{diag}(s_y,\ldots,s_y, s_u,\ldots,s_u) \tag{2.4.13}$$

so C has rank 2n a.s. If the stochastic part is just independent plant noise, as in Theorem 1, then

$$P = \text{diag}(s,\ldots,s,0,\ldots,0)$$

so C is still invertible if $2n \leq 2m+n$, and Theorem 1 is still valid. The small sample properties of Theorem 3 are also correct if $n \leq 2m$, but the full matrix formulation is also unbiased if $n > 2m$.

It is not clear exactly what happens asymptotically (when the number of samples becomes large). The estimates do converge to some estimate; if there is just independent plant noise, the estimate is an exact realization of the system. It would be very difficult to decide which realization is approached, because it would require a very detailed probabilistic argument on the errors in (2.1.3) and knowledge of the relation between $(D+P)^{-1}$ and P^{-1} and D^{-1} in (2.4.12) which not available. It would be nice to say that the answer was either the noise-free least squares estimate or realization (2.4.3). In fact numerical simulations in Chapter III and in (A2) have sometimes indicated one and sometimes the other of these. The reason for this lies in the fact that we are trying to minimize the norm of the parameter vector when there is more than one realization. Figure 1 shows the norm of the estimate for Example 1 when $n = 4$; while there is a minimum at the lse, the curve is very flat. Soderstrom (S1) indicates that machine roundoff error alone can cause a large error in the estimate, so it is not surprising that simulations have not generated consistent results.

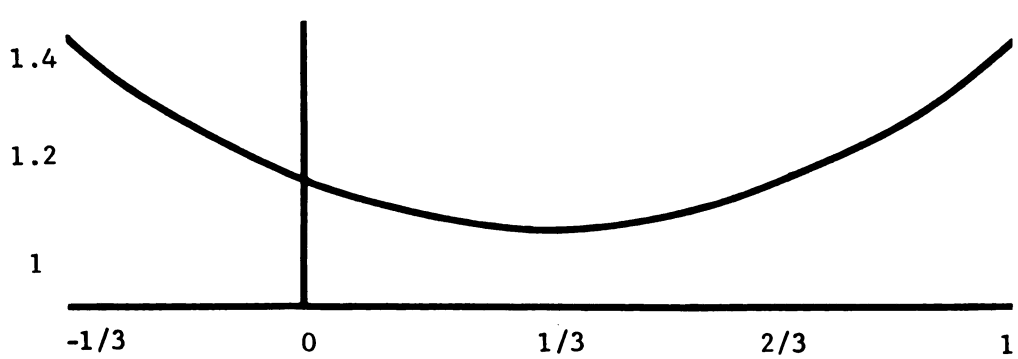

Figure 1. Norm of θ vs. Extra Pole

If we cannot say much theoretically about the asymptotic estimate, we can say even less about the small sample estimates. The natural approach is to let P, the stochastic part of $E(C_N)$ approach zero, but then the rank of $E(C_N)$ changes from 2n to, at most, (n+m). When the rank of a matrix changes, the (pseudo-) inverse is no longer a continuous function of its elements as the following examples show.

$$\begin{bmatrix} e & 0 \\ 0 & e \end{bmatrix}^{-1} = \begin{bmatrix} 1/e & 0 \\ 0 & 1/e \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}^{+} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1+e & 1 \\ 1 & 1+e \end{bmatrix}^{-1} = \frac{1}{e^2+2e} \begin{bmatrix} 1+e & -1 \\ -1 & 1+e \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}^{+} = \frac{1}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Thus, the noisy lse might, theoretically, be quite different from the no-noise lse for small samples.

But the numerical results in Chapter III are much different than the pessimistic theory. Even with noise the computed poles and zeros are often quite close to the noise-free values.

2.4.3 Tests for the order of the system

The literature contains very few papers on the model-order problem per se (A2, G2, W3), but there has been work on choosing a regression model from a set of possible models. All of the existing work on choosing the order of a linear, dynamic system has just modified the general regression tests to fit the particular problem. These existing tests fit into two categories. We can say that ϕ has full rank a.s. if and only if $n \leq m$ and that the problem is to determine the largest n so that the columns

of ϕ are sufficiently independent. On the other hand, we can look at the residuals and say that they should have zero mean and approximately a Gaussian distribution if the correct order is chosen. None of these methods make use of the considerable body of knowledge of linear systems that is available. After a rather complete review of the existing tests of model order, with brief indications of the numerical results in Chapter III, we shall present some new tests which do make use of the theory of linear systems, in particular the theory of equivalent representations.

The second point of view, residuals, is more common in the engineering literature (A2, J1, W3). There are several non-rigorous tests based on the residuals. Draper and Smith (D1) outline several ways to examine the residuals of a given model such as plots of the residual versus time, the input, and the output, examination of outliers (those residuals with large magnitudes), and run tests; none of them seem very useful except at a very preliminary stage of investigation. A much better one, originally proposed by Forsythe (F2) and since used by several others (B4, J1, W3), is often useful. Forsythe considers the problem of determining the degree of a polynomial when only noisy values are available and gives the following heuristic algorithm: Add one degree at a time to the approximating polynomial; the sum of the squared residuals (or the RMS residual) should decrease rapidly until the correct degree is reached and then remain fairly constant. Boling (B4) has modified the algorithm to include a stopping rule based on the relative decrease. Numerical examples indicate that Forsythe's method works well in determining the order of a dynamic

system when the variables have a wide enough range but does not work well when the output is relatively constant. Other residual norms, such as the largest absolute one, give similar results.

In all the rigorous tests based on the residuals, it is assumed that they are independent samples from a zero-mean Gaussian distribution. This restriction may be more important theoretically than practically, at least in part because of the central limit theorems. The most common of these is an F-test (A1, E1, B1). Let V_1 and V_2 be the sum of the squared residuals for models of order n_1 and $n_2 > n_1$, respectively, over the same N observations. Then

$$F = \frac{V_1 - V_2}{V_2} \cdot \frac{N - 2n_2}{2(n_2 - n_1)} \tag{2.4.14}$$

has an F-distribution with $2(n_2 - n_1)$ and $(N - n_2 - 1)$ degrees of freedom. The F-test says to accept the hypothesis that

$$n_2 > n_1 \geq m$$

if F is smaller than an appropriate value of the F-distribution and reject the hypothesis otherwise. This test also works well when the variables have a wide enough range. Another test on the residuals (J1) is to compute the sample spectrum of the residuals and test the integrated spectrum using a Kolmogorov-Smirnov statistic.

The first point of view, the rank of ϕ, is more common in the econometrics literature (F1, H1, W2), where it is called the multicollinearity problem (the word identification being reserved for a different problem). ϕ has full column rank if and

only if $\phi^T\phi$, a square matrix, is invertible, so tests based on $\det(\phi^T\phi)$ seem reasonable. Both Farrar and Glauber (F1) and Haitovsky (H1) do so. However, they use the correlation matrix, whose determinant is between zero and one and is based on a zero-mean process, instead of the unstandardized $\phi^T\phi$. Haitovsky's test seems more reasonable: Let R_n be the $2n \times 2n$ correlation matrix. Then, as in (H1),

$$X = -(N + 2n/3 + 11/6)\ \ln(1 - \det R_n) \tag{2.4.15}$$

has a chi-squared distribution with $n(2n-1)$ degrees of freedom. A small X value implies that $n > m$. For a dynamic system, it is easy to show that $\det R_{n+1} \leq \det R_n$ so that the test gives a stopping criterion. Of course, if the input or output is relatively constant, then R_n will be small for any $n > 1$. The other problem is that by subtracting the means we may have changed the rank of the matrix under investigation, i.e. R_n and $\phi^T\phi$ may not have the same rank. $\phi^T\phi$ can be normalized so that its determinant is between zero and one, but the chi-squared test will still not apply. Tests based on $\det(R_n)$ or $\det(\phi^T\phi)$ do not seem to work as well as the residual tests.

Goodman and Hiller (G2) give a test based on the determinant of a square ϕ and estimates of the maximum errors. But, again, the determinants will be small if a variable is almost constant and will be very sensitive to the particular values of the variables. Some sort of average determinant might be useful to reduce this dependency, but the algorithm is not too promising.

As mentioned above, all of these tests apply to the general problem of choosing a regression model. The only applications to dynamic systems have been the Goodman and Hiller paper (G2), an application of the F-test by Astrom (A2), and a paper by Woodside (W3). Woodside compares three tests: the mean squared residual, the ratio of the determinants of unnormalized $\phi^T\phi$ matrices, and a likelihood ratio test which is fairly complicated computationally. He concludes that the mean squared residual test is the best of these. He also gives a method for improving the results if the characteristics of the (uncorrelated) noise are available.

All of the above tests have the disadvantage that they do not work well in practice when the input or output is almost constant. This is expected in the general regression situation. But we can expect better results if we can use knowledge of dynamic systems. Our basic problem is to find a model which behaves in the same manner as the actual system, at least in those aspects of interest. Ideally, the model might have the same order, coefficients, poles, and zeros as the (minimal respresentation of the) system, but this might not be possible or even necessary. If we have several models which behave similarly to each other, then any of them should be acceptable. In particular, we can use knowledge of equivalent representations and reduction of dynamic systems.

Our procedure will have four steps. First, we compute the parameters for a number of models; for example, we may get models of orders one through M for the full matrix, reduced matrix,

and normalized least squares estimators (more will be said about the selection of the algorithm in Chapter III). It might be possible to select the order on the basis of the mean squared residual or other such test. If not, these tests will allow us to eliminate some possibilities. This is the second step and reduces our selection space to, say, orders m' through M'. The third step is to examine the poles and zeros of the estimated transfer functions. The numerator and denominator of an estimate should have approximately common factors if its order is larger than the minimal order of the system. We can also look at the poles and zeros of different models; they should be approximately constant with respect to changes of model order. The theoretical problems with the third step are that the roots of a polynomial are notoriously sensitive to small changes or errors in the coefficients, and that the standard error of the coefficients increases as the order does and as the variables become more constant. However, the numerical results are very good. The final step is to choose the simplest acceptable model based on steps two and three.

There is a test based on the theory of dynamic systems which has shown better numerical results than any of the others. In Section 1.2.1, we stated that, given the controllability matrix, Q_c, and the observability matrix, Q_o, of any realization of a system, the minimal order of the system is $\text{rank}(Q_o^T Q_c)$. Since the observability matrix of realization (1.2.3)-(1.2.6) is the identity, the minimal order is $\text{rank}(Q_c)$. Since the stochastic elements will cause Q_c to always have maximum rank, it seems reasonable to compute $\det(Q_c)$ and decide when the determinant becomes small

enough as the rank changes.

It is easy to set up Q_c and compute its determinant because it is a Hankel matrix, i.e.

$$Q_c(i,j) = s_{i+j-1} \tag{2.4.16}$$

To see this recall that

$$Q_c = (B,\ AB,\ldots,A^{n-1}B) \tag{2.4.17}$$

It is now easy to see that

$$s_j = h_j,\ j = 1,2,\ldots,n \tag{2.4.18}$$

where h_j is defined in (1.2.3), and a little computation shows that

$$s_j = \Sigma_1^n s_{j-i} a_i,\ j = (n+1),\ldots,(2n-1) \tag{2.4.19}$$

2.5 Normalized Least Squares

One problem with least squares is that larger data values influence the results more than smaller values. For example, let

$$y_t = y_{t-1} - u_{t-1} \tag{2.5.1}$$

We will take two sequences of observations and find the least squares solutions; in each sequence there is just one error, y_4.

y_1	u_1	y_2	u_2	y_3	u_3	y_4		y_1	u_1	y_2	u_2	y_3	u_3	y_4	u_4
1	1	0	1	-1	10	-12		10	10	0	10	-10	1	-12	

In matrix form

$$\begin{bmatrix} 0 \\ -1 \\ -12 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ -1 & 10 \end{bmatrix} \theta; \quad \begin{bmatrix} 0 \\ -10 \\ -12 \end{bmatrix} = \begin{bmatrix} 10 & 10 \\ 0 & 10 \\ -10 & 1 \end{bmatrix} \theta$$

Solving for θ

$$\theta = \begin{bmatrix} 135 \\ -134 \end{bmatrix} \frac{1}{123} \qquad \theta = \begin{bmatrix} 342 \\ -332 \end{bmatrix} \frac{1}{321}$$

and the errors are

$$E = \frac{1}{123} \begin{bmatrix} -1 \\ 11 \\ -1 \end{bmatrix} \qquad E = \frac{1}{321} \begin{bmatrix} -10 \\ 110 \\ 542 \end{bmatrix}$$

Thus, in the first sequence, the last residual is the smallest even though it is the only line where there is an error. In the second sequence, the last residual is the largest.

Werther (W1) has proposed a weighting scheme which eliminates this inequity. He divides the elements in each row by the rms value of the row; this corresponds to premultiplying by a diagonal weighting matrix. He states that the results are improved sometimes by this "normalization" process.

The non-uniform effect of the observation magnitudes is also important if the input and output have different magnitudes. In Examples 2 and 3, Chapter III, the output is about twice the input. Not only does the output have more influence on the parameter estimates, but the variances of the input coefficients are larger than the variances of the output coefficients. This is because the variance of a coefficient equals the appropriate diagonal term of $(\phi^T \phi)^{-1}$. For a first order system, the variance

of the output coefficient is proportional to the sum of squares of the input, and the variance of the input coefficient is proportional to the sum of squares of the output. In the examples the input coefficient has four times the variance of the output coefficient even though it is three orders of magnitude smaller!

Attempts we have made to correct this problem have not been too successful. More work needs to be done on the area of choosing a good weighting matrix.

2.6 Summary

This chapter contains the theoretical results on least squares and the model order problem. After explicitly defining the full and reduced matrix least squares algorithms, it was shown that they converge in mean square with correlated errors. The small sample bias of the full matrix algorithm and unbiasedness of the reduced one were shown. Several tests for the order of the system were given in Section 2.4.3 including some new ones based on systems realization theory. The last section contained some material on the selection of weighting schemes to normalize the effect of the observations.

CHAPTER III

NUMERICAL COMPUTATIONS

In this chapter we will compare the algorithms and order tests defined in Chapter II. First, we will describe the numerical algorithms used in the computations. Then the examples will be described in some detail. Finally, the numerical comparisons will be made using both simulated data and observed data from a real, but unknown, system.

3.1 Description of the Algorithms Used

There are several ways to compute the least-squares estimate off-line. The most obvious, and the most common, method is to pre-multiply both sides of (2.1.3) by ϕ^T and solve the system of equations

$$(\phi^T\phi)\theta = \phi^T Y \tag{3.1.1}$$

for θ by the Gaussian algorithm. But this method is numerically unstable (C3). A much better method is to determine a unitary matrix Q such that $Q\phi$ is triangular. Then the solution of

$$(Q\phi)\theta = QY \tag{3.1.2}$$

is the same as the solution of (3.1.1) and the least squares solution of (2.1.3), but it is much less prone to numerical errors.

This algorithm is the one which we have used most often because of its stable numerical properties; the actual computer subroutine was obtained from the IBM Scientific Subroutine Package (I1).

Least squares can also be done on-line as follows (G3): write the n-th order difference equation representation as

$$y_{n+i} = \theta^T z_i \tag{3.1.3}$$

where

$$z_i = \mathrm{col}(y_i,\ldots,y_{n-1+i},\ u_i,\ldots,u_{n-1+i}) \tag{3.1.4}$$

and let

$$U_N = \Sigma_1^N\ y_{n+i} z_i^T$$

$$T_N = \Sigma_1^N\ z_i z_i^T$$

Then

$$\theta_N = U_N T_N^{-1} \tag{3.1.5}$$

This algorithm can be easily modified to handle multi-input multi-output systems (G3). It also requires less storage than the off-line algorithms since it is necessary to keep only the latest U_N vector and the latest T_N matrix, in addition to one z_i vector. However, some computational effort can be saved at the expense of a little bit of storage by finding the inverses recursively. To do this, we divide the algorithm into three stages. T_N is invertible a.s. if $N \geq 2n$, so we can use the well known (G3, P1) equation of (3.1.7) to compute T_{N+1}^{-1} from T_N^{-1}. If $N < 2n$, we use the first two stages to get T_{N+1}^{+} from T_N^{+}.

Step (a) $N = 1$. Then

$$T_1^+ = \left(\frac{1}{T_{z_1 z_1}}\right)^2 T_1 \tag{3.1.6}$$

Proof: By definition, $T^{\#} = T^+$ if $TT^{\#}T = T$, $T^{\#}TT^{\#} = T^{\#}$, and $TT^{\#}$ and $T^{\#}T$ are hermitian. Dropping the subscript, $T^2 = zz^T zz^T = z\lambda z^T = \lambda T$, where $\lambda = z^T z$. Thus, $TT^+T = (1/\lambda)^2 T^3 = (1/\lambda)^2 \lambda T^2 = T$. Similarly, $T^+TT^+ = T^+$. TT^+ and T^+T are hermitian because T is.

Steb (b) For $1 < N < 2n$, when T_N is not invertible, then we can use the new algorithm given below.

Step (c) If T_N is nonsingular, then

$$T_{N+1}^{-1} = T_N^{-1} - \frac{1}{1+z_{N+1}^T (T_N^{-1} z_{N+1})} (T_N^{-1} z_{N+1})(T_N^{-1} z_{N+1})^T \tag{3.1.7}$$

The algorithm for step (b) requires that we write T_N as

$$T_N = \Sigma_1^N z_i z_i^T = ZZ^T \tag{3.1.8}$$

where

$$Z = (z_1, \ldots, z_N) \tag{3.1.9}$$

Then, if $W = Z^T Z$ is nonsingular

$$T_N^+ = ZW^{-2}Z^T \tag{3.1.10}$$

Proof: $TT^+ = ZZ^T ZW^{-2}Z^T = ZWW^{-2}Z^T = ZW^{-1}Z^T$. Thus, $TT^+T = ZW^{-1}Z^T ZZ^T = ZWW^{-1}Z^T = ZZ^T = T$. The rest of the proof is similar.

While the algorithm holds when the z_i are matrices, that is in the multi-input, multi-output case, if they are vectors we do not have to invert W at each step. First note that W_{N+1} and W_N differ only by the addition of a new column and row.

$$W_{N+1} = \begin{bmatrix} W_N & b \\ b^T & c \end{bmatrix} \tag{3.1.12}$$

where b is $N \times 1$ and c is a scalar. Thus (F3)

$$W_{N+1}^{-1} = \frac{1}{d} \begin{bmatrix} dW_N^{-1} + ee^T & -e \\ -e^T & 1 \end{bmatrix} \tag{3.1.13}$$

where

$$\begin{aligned} d &= c - b^T W_N^{-1} b \\ e &= W_N^{-1} b \end{aligned} \tag{3.1.14}$$

W_{N+1} is nonsingular if W_N is and d is non-zero.

Another on-line method is stochastic approximation (M3).

$$\theta_{N+1} = \theta_N + P_{N+1} z_N (y_{n+N} - z_N^T \theta_N) \tag{3.1.15}$$

where

$$P_{N+1} = P_N - \frac{1}{1 + z_N^T P_N z_N} (P_N z_N)(P_N z_N)^T \tag{3.1.16}$$

The initial matrix P_0 may be any positive definite symmetric matrix of appropriate dimensions, but is usually taken as a multiple of the identity. The stochastic approximation estimates approach the least squares estimates as N becomes large.

Stochastic approximation has the advantage over on-line least squares that it is self-correcting; the factor $(y_{n+N} - z_N^T \theta_N)$ is the error of the last observation. It will also automatically correct for slowly-varying parameters. Accumulated roundoff error was a problem for on-line least squares when the stiff example, Example 2, was used; double precision arithmetic corrected this problem. While the on-line algorithms are useful for control of an operating system, the off-line algorithm seems preferable for the model order problem because it allows the least amount of overall error.

The determinants were taken by the pivotal condensation method (F3). Some reduction in computation could have been saved by using elementary row and column operations on the original matrices, but the necessary bookkeeping outweighed the improvement. Both the correlation and observability matrices have determinants that are relatively insensitive to round-off errors in most situations.

3.2 Description of the Examples

In the next section, we will use three examples to numerically compare the algorithms and tests. The first is regular simulated system; the second an ill-conditioned simulated system; and the third is actual data observed on an electric power network.

Example 1. Regular system

$$y_t = .8y_{t-1} + .39y_{t-2} - .27y_{t-3} - .5u_{t-1} + .5u_{t-2} + .1u_{t-3} \tag{3.2.1}$$

To illustrate the procedure of Section 2.4.1 for the calculation of the extra zeros when $n > m$, we will do it for $p = n-m = 1$ and for $p = 2$. For $p = 1$, (2.4.6) and (2.4.7) become

$$\theta_1^T = (-1, 0, .8, -.5, .39, .5, -.27, .1)$$

$$\theta_0^T = (.8, -.5, .39, .5, -.27, .1, 0, 0) \tag{3.2.2}$$

$$H = \theta_1^T$$

Thus

$$G = -(\theta_1^T\theta_1)^{-1}\theta_1^T\theta_0 = -\frac{-.7933}{2.375} \approx 1/3 \tag{3.2.3}$$

The additional pole of $T(z)$ is about -1/3.

For $p = 2$ we have

$$\theta_2^T = (0, 0, -1, 0, .8, -.5, .39, .5, -.27, .1)$$

$$\theta_1^T = (-1, 0, .8, -.5, .39, .5, -.27, .1, 0, 0)$$

$$\theta_0^T = (.8, -.5, .39, .5, -.27, .1, 0, 0, 0, 0)$$

Now

$$G = -\begin{bmatrix} 2.375 & -.7933 \\ -.7933 & 2.375 \end{bmatrix}^{-1} \begin{bmatrix} -.7933 \\ -.656 \end{bmatrix} = \begin{bmatrix} .47982 \\ .43648 \end{bmatrix} \tag{3.2.4}$$

Thus, the additional factor is $(z^2 + .47982z + .43648)$ and the additional poles are $(-.2399 \pm .6156i)$.

Example 2. Ill-conditioned system

$$\begin{aligned} y_t = {} & .1998y_{t-1} + .3998y_{t-2} + .20792y_{t-3} \\ & + .1035616y_{t-4} + .0883232y_{t-5} - .55E\text{-}4u_{t-1} \\ & + .1595E\text{-}3u_{t-2} - .14245E\text{-}3u_{t-3} + .45925E\text{-}4u_{t-4} \\ & + .88195E\text{-}3u_{t-5} \end{aligned} \tag{3.2.7}$$

The poles of the transfer function are 0.9998, (-.5 $\pm$ .3i), and (.1 $\pm$ .5i); the zeros are approximately 2.8079, -1.34194, and (.717 $\pm$ 1.934i). This fifth-order example was designed to approximate the power system of Example 3; an average input of 33.4 should yield an average output of 60.0. The input used in the simulations was approximately uniformly distributed in the interval (30.4, 36.4); the output had a mean of about 59.4 and a standard deviation of about .31. The range of the output, although much larger than the output of example, was relatively small, about .25% of the mean.

Example 3. Observed data

The last example is a set of load-frequency data taken from an actual operating electric power system. An attempt to model the data was made using stochastic approximation (P1); several "good" models were made, and the problem of selecting the best of them motivated the present work.

3.3 Computed Results

In this section we examine the output of simulated runs of examples one and two, where the noise has different properties on each run, and the estimates for the real data of example three. We will see that the model tests alone are sufficient to identify the order of the regular system, Example 1, but are not sufficient for the power example and its approximate simulation, Example 2.

There was no difficulty choosing the order of Example 1. With even a small number of sample points, N = 100, the controllability matrix test worked with every noise configuration that

was used; the mean squared residual worked except at high noise levels; the F-test except when there was only observation noise; the correlation matrix, and its chi-squared variant, did not work. There seemed to be little difference between the full and reduced matrix results.

Figures 2, 3, 4 and 5 and Tables 2, 4 and 5 compare the results for three different simulations of Example 1. All three had relatively high noise levels (one had system noise only, one observation noise only, and one had both), and the full matrix tests were more definite than the reduced ones for all three. Figure 2 shows the absolute values of the controllability determinants; the numerical data for two of the runs is in Table 2. As shown in Figure 3, there is at least an order of magnitude change when going from $n = 3$, the correct order, to $n = 4$, except in two instances, and the drop was often two orders of magnitude. In the two exceptions, the large drop occurred from $n = 4$ to $n = 5$, giving an estimated order of four.

Figure 4 contains a graph of the full matrix mean squared residuals, listed in Table 2. The three runs shown there were the only three where this test did not work well (more typical runs are in Figure 5); while there is very little decline after $n = 3$, there is also very little from $n = 2$ to $n = 3$. This test gives an estimated system order of two. Figure 4 also shows the determinants of the correlation matrices for one of the runs; the straight line behavior was typical, indicating that the test is not very good. The chi-squared test based on this determinant was almost a complete failure.

Table 2. Order Tests for Example 1

order	Mean Square Residual full	Reduced	Det(Q) Full	Reduced
1	.0794	.0824	.0846	.1103
2	.0277	.0235	.0660	.1390
3	.0247	.0216	.0401	.1106
4	.0245	.0213	.0007	.0262
5	.0244	.0212	.0001	.0036
	System and observation noise			
1	.0770	.0790	.0899	.1062
2	.0528	.0215	.0799	.1194
3	.0224	.0194	.0462	.0944
4	.0223	.0189	.0006	.0158
5	.0221	.0187	.0000	.0003
	System noise only			
1	.0498	.0480	.1871	.2461
2	.0041	.0030	.0967	.1722
3	.0020	.0017	.0312	.0456
4	.0017	.0014	.0015	.0003
5	.0014	.0011	.0003	.0002
	Observation noise only			

$$y_t = .8y_{t-1} + .39y_{t-2} - .27y_{t-3} - .5u_{t-1} + .5u_{t-2} + .1u_{t-3}$$

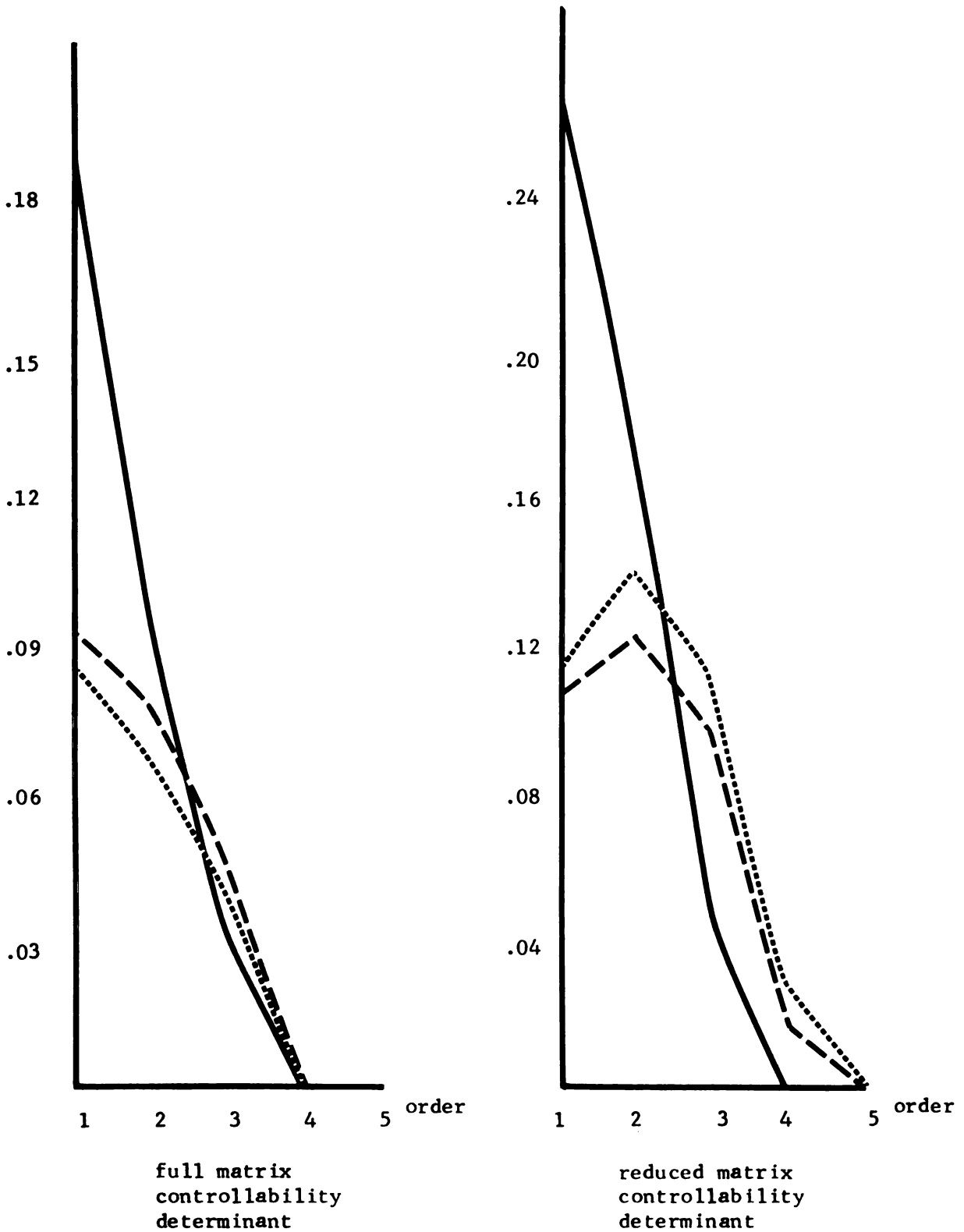

Figure 2. Determinants of Controllability Matrices.

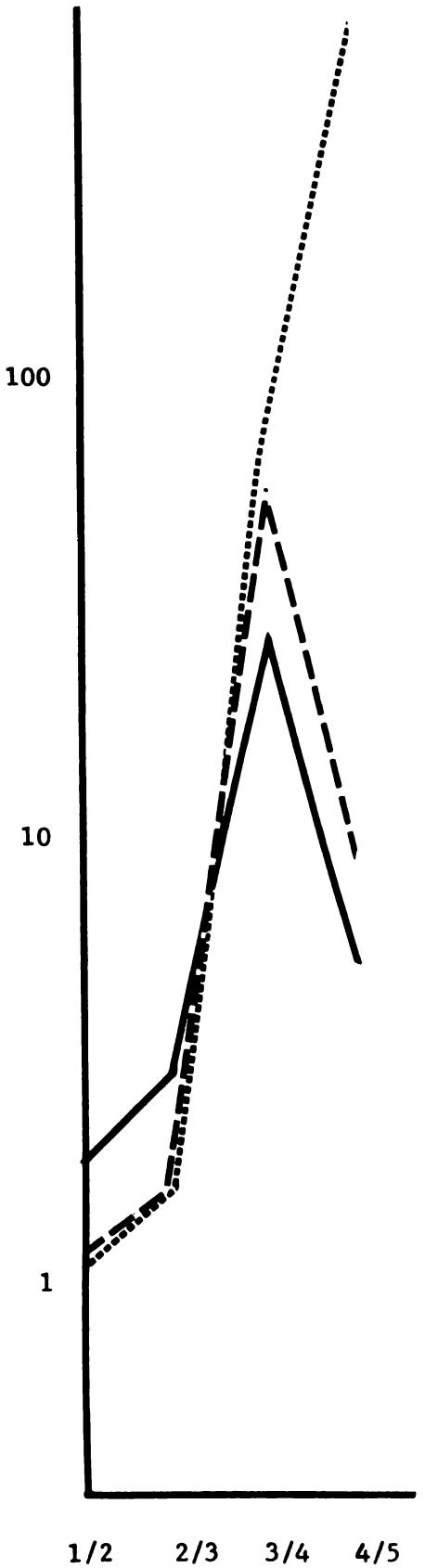

Figure 3. Ratio of Controllability Determinants
(full matrix lse)

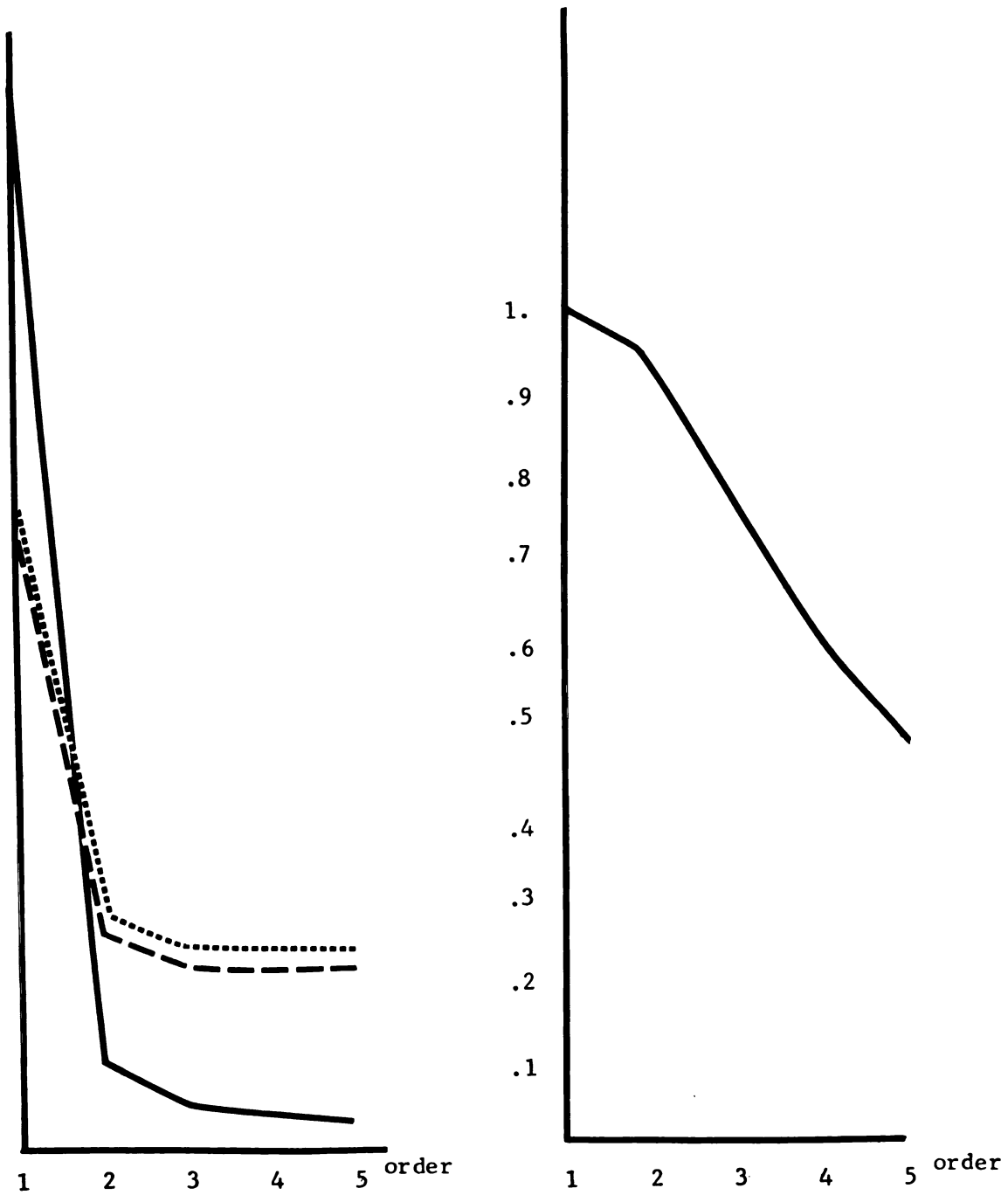

Figure 4. Order Tests

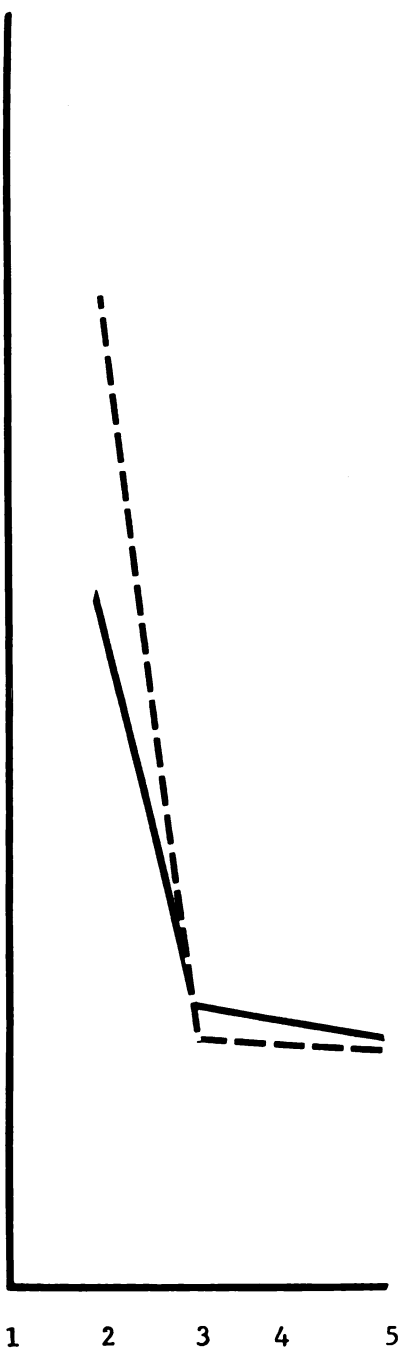

Figure 5. Mean Square Residuals

Table 4 shows the F-test arrays for these runs. The critical values at 90% confidence are in Table 3.

Table 3. Critical Values for the F-test

$n_2 - n_1$	1	2	3	4
value	2.30	1.94	1.77	1.67

If a value in the array is lower than the one in Table 3, then n_2, $n_1 \geq m$, the minimal order. The first pair of arrays in Table 4 are for observation noise only, and both indicate that the system order is at least five. The test does work for the other two noise configurations shown, although the last array does not have completely consistent results.

Table 5 summarizes the results of the test on the noise configurations shown for Example 1. We can conclude that, even for these relatively high noise levels, the order of the system is three and we need not do anything further. We can also conclude that the determinant of the controllability matrix is the test that works most often.

The problem of choosing the system order is more difficult for the second example, as the summary Table 6 shows. The F-test, detailed in Table 7, is not very helpful. The mean squared residual, Figure 6 and Table 8, becomes flatter as n increases, but there is no one point where there is a large change of slope as there was in Figure 5. But the controllability determinant does work in most cases.

Table 4. F-test for Example 1

Order	1	2	3	4
2	2745.			
3	2929.	259		
4	2373.	182.	52.0	
5	2148.	162.	56.2	50.1
	Full matrix -- observation noise only			
2	732.			
3	635.	34.1		
4	528.	27.2	12.1	
5	487.	25.9	13.0	11.2
	Reduced matrix -- observation noise only			
2	504.			
3	301.	33.0		
4	201.	17.0	1.11	
5	152.	11.9	1.37	1.62
	Full matrix -- system noise only			
2	129.			
3	7.21	4.90		
4	48.7	3.06	1.20	
5	36.3	2.22	.886	.584
	Reduced matrix -- system noise only			
2	464.			
3	274.	30.0		
4	184.	15.9	1.62	
5	138.	11.0	1.46	1.30
	Full matrix -- system and observation noise			
2	120.			
3	66.0	4.11		
4	43.9	2.37	.652	
5	32.4	1.62	.429	.218
	Reduced matrix -- system and observation noise			

Table 5. Estimates of the System Order

Noise type	Full MSR	Reduced MSR	Full F-test	Reduced F-test	Full det(Q)	Reduced det(Q)
System	2 or 3	2 or 3	3	3	3	3 or 4
Obser.	2 or 3	2 or 3	≥ 5	≥ 5	3	3
Both	2 or 3	2 or 3	3	3	3	3

Table 6. Order Test Summary for Example 2

Test	Observation noise		System noise		Both	
	Full	Reduced	Full	Reduced	Full	Reduced
MSR	X	5 or 7	5	6	4 or 5	X
F-test	≥ 7	4	5 or 6	3 or 5	5	4 or 5
$\mathrm{Det}(Q_c)$	5	4 or 5	5	5	5	5
$\frac{\mathrm{Det}(Q_c)}{\pi b_i}$	5	5	5	5	5	5

Note that the determinant (Table 8) decreases by about three orders of magnitude in almost all situations and that the relative decrease is approximately equal to the input coefficients, b_i. This observation led to a heuristic modification which has given very satisfactory results in every instance of Examples 1 and 2: "normalize" the determinant by dividing by the product of the b_i. Then

$$\frac{\det(Q_{m-1})}{\pi_1^{m-1} b_{i,m-1}} \ll \frac{\det(Q_m)}{\pi_1^{m} b_{i,m}} \tag{3.3.1}$$

Table 7. F-tests for Example 2

order	1	2	3	4	5	6
2	91.1					
3	78.0	50.0				
4	62.8	37.5	21.6			
5	53.4	31.6	19.3	16.0		
6	45.2	26.0	15.6	11.8	7.29	
7	42.1	24.9	16.1	13.4	11.6	15.5
	Full matrix -- observation noise only					
2	14.2					
3	9.76	4.06				
4	8.42	4.22	4.02			
5	7.09	3.65	3.19	2.21		
6	5.60	2.75	2.17	1.22	.275	
7	5.02	2.56	2.06	1.36	.938	1.60
	Reduced matrix -- observation noise only					
2	138.					
3	102.	46.3				
4	78.6	33.9	18.7			
5	65.2	28.3	16.9	14.2		
6	52.4	21.6	11.7	7.77	1.30	
7	44.2	17.7	9.28	5.84	1.61	1.93
	Full matrix -- system and observation noise					
2	16.9					
3	11.5	4.47				
4	8.80	3.52	2.39			
5	7.37	3.15	2.33	2.16		
6	6.43	2.90	2.22	2.05	1.89	
7	5.22	2.27	1.63	1.36	.958	.084
	Reduced matrix -- system and observation noise					
2	156.					
3	99.8	28.7				
4	70.2	18.2	6.98			
5	63.4	21.6	16.6	25.6		
6	51.7	17.1	12.2	14.4	3.11	
7	43.0	13.7	9.15	9.67	1.64	.181
	Full matrix -- system noise only					
2	14.9					
3	9.76	3.55				
4	6.55	1.97	.449			
5	5.81	2.25	1.55	2.62		
6	4.93	1.98	1.41	1.88	1.14	
7	3.99	1.57	1.07	1.27	.628	.152
	Reduced matrix -- system noise only					

Table 8. Order Tests for Example 2

order	Mean Sq. Residual		Det(Q)		Normalized Det(Q)	
	Full	Reduced	Full	Reduced	Full	Reduced
1	.773E-4	.778E-4	.231E-3	.158E-3	1.0	1.0
2	.554E-4	.555E-4	.498E-7	.114E-6	.703	1.422
3	.474E-4	.497E-4	.605E-10	.186E-10	2.36	.468
4	.442E-4	.444E-4	.661E-14	.236E-12	2.71	6.83
5	.419E-4	.415E-4	.514E-15	.187E-14	443.	134.
6	.409E-4	.412E-4	.159E-18	.118E-17	2941.	264.
7	.388E-4	.391E-4	.909E-22	.127E-20	3425.	887.
		Observation noise only				
1	.136E-3	.120E-3	.138E-3	.145E-3	1.0	
2	.927E-4	.812E-4	.241E-6	.475E-7	1.65	
3	.802E-4	.719E-4	.223E-9	.171E-9	4.06	
4	.757E-4	.671E-4	.161E-14	.629E-13	.502	
5	.720E-4	.629E-4	.103E-14	.319E-14	654.	
6	.717E-4	.594E-4	.182E-18	.482E-18	366.	
7	.712E-4	.592E-4	.921E-21	.303E-21	8033.	
		System and observation noise				
1	.649E-4	.575E-4	.110E-3	.152E-3		
2	.426E-4	.405E-4	.140E-6	.147E-7		
3	.388E-4	.367E-4	.613E-10	.143E-10		
4	.379E-4	.362E-4	.677E-14	.301E-14		
5	.349E-4	.335E-4	.106E-14	.269E-15		
6	.345E-4	.323E-4	.352E-18	.861E-20		
7	.345E-4	.322E-4	.523E-21	.614E-24		

System noise only

$$y_t = .1998y_{t-1} + .3998y_{t-2} + .2079y_{t-3} + .1035616y_{t-4} + .0883232y_{t-5}$$

$$-.55E\text{-}4u_{t-1} + .1595E\text{-}3u_{t-2} - .14245E\text{-}3u_{t-3} + .45925E\text{-}4u_{t-4} + .88195E\text{-}3u_{t-5}$$

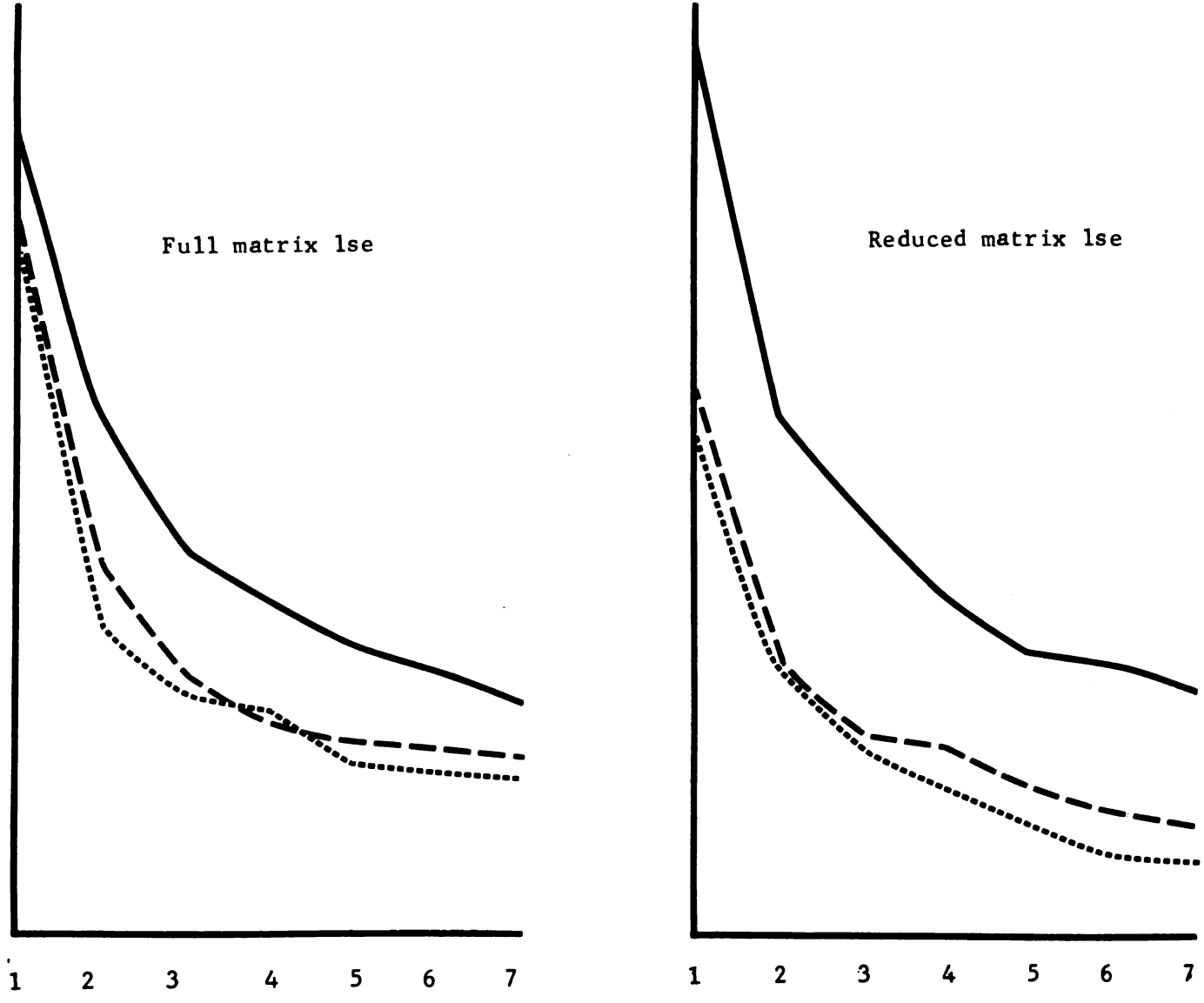

Figure 6. Mean Squared Residual for Example 2

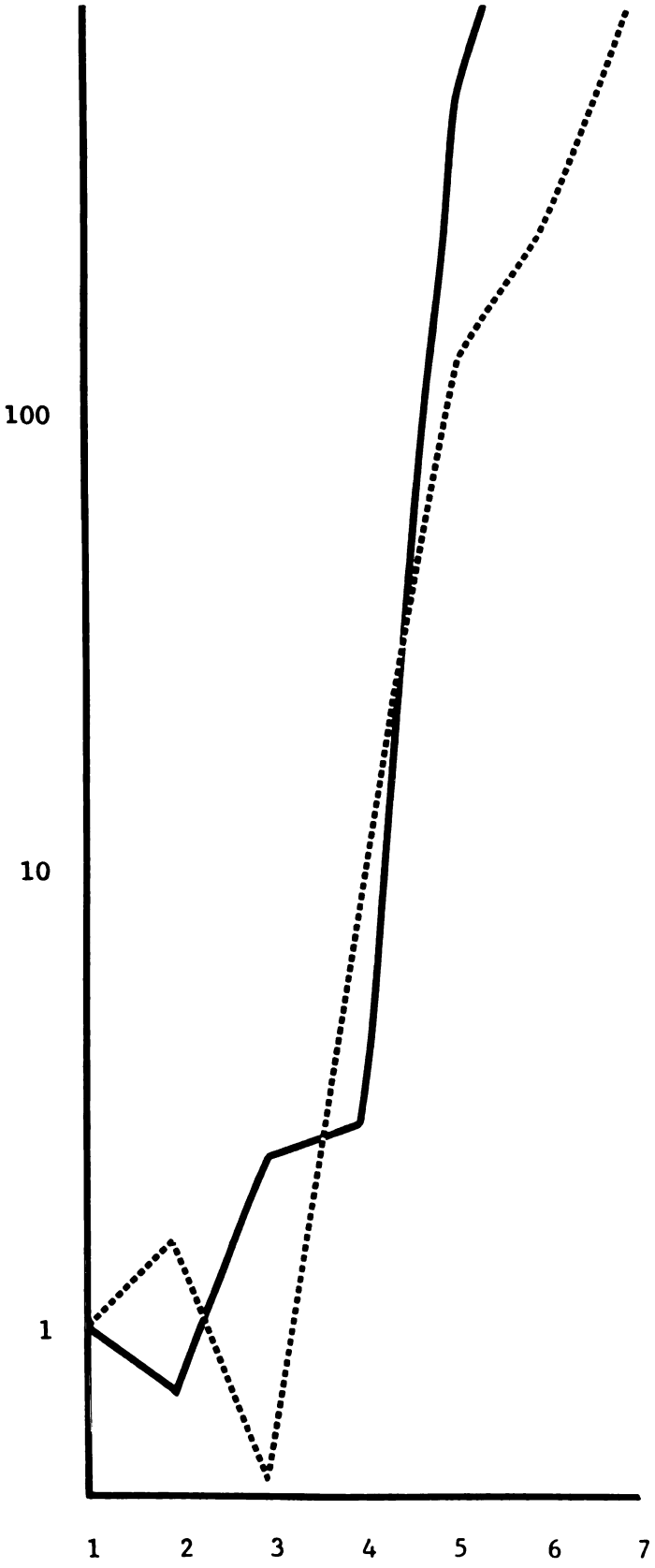

Figure 7. Normalized Controllability Determinant for Example 2

where the subscript denotes the order of the model, and m is the minimal order of the system. Table 8 also shows these normalized determinants. On the basis of them, we can conclude that the minimal system order is five, the correct one, in every instance.

If we are not satisfied with the conclusion from the controllability test, then we can look at the poles and zeros of the transfer functions. Table 9 shows one set of poles and zeros. There is very good agreement, with respect to change of model order, on the dominant pole. Starting with the third order there is a complex pair with negative real part and norm around .6; these will behave very similarly. Another pair with about the same norm enters at $n = 5$. Looking at the zeros, we see that starting with $n = 5$ there are two real (around -1.3 and -3) and one complex pair (with positive real part and norm 1.1). While the fourth-order poles are similar to the fifth-order ones, the fourth and fifth order zeros are not similar. There is no approximate pole-zero cancellation to help decide. But we can get some information from the steady-state response to a unit-step input, also shown in Table 9. There is very good agreement starting with $n = 5$, and less agreement with lower order values. Thus we can conclude from examination of the transfer functions, as well as from the normalized controllability test, that the system is fifth order.

The order tests yield more consistent answers for Example 3.

Table 9. Poles and Zeros for Example 2

(a,b) denotes the complex pair $(a \pm bi)$

order	Poles			
2	.99969			-.561
3	.99978	(-.380, .466)		
4	.99975	(-.120, .622)		-.607
5	.99962	(-.498, .284)	(.057, .627)	
6	.99967	(-.322, .309)	(.105, .588)	-.442
7	.99965	(-.629, .289)	(.322, .489)	(-.131, .661)
True	.9998	(-.5, .3)	(.1, .5)	

order	Zeros				Steady state
2	-1.38				1.653
3		-2.21		.315	1.600
4		-2.24	(.106, .435)		1.621
5	-1.13	-3.16	(.591, .915)		1.675
6	-1.29	-2.81	(.504, .941)	.285	1.661
7	-1.31	-2.79	(.543, .881)	(.134, .382)	1.665
True	-1.34	+2.81	(.717, 1.93)		1.603

Table 10. Order Test Summary for Example 3

Test	Full lse	Reduced lse
Mean Squared Residual	X	4
Controllability det	4	4
Normal det(Q_c)	4 or 6	4 or 6
F-test	5	4

The tests are detailed in Tables 11 and 12 and in Figures 8 and 9. From them we draw the tentative conclusion that the system can be approximated fairly well by a fourth order model.

The poles and zeros for both the full and reduced algorithms are in Table 13. As in Example 2, the poles and zeros are relatively constant as the model order changes, but there is no pole-zero cancellation. The steady-state response to the unit-step input is virtually constant for the full lse and, as expected, varies slightly for the reduced lse. The examination of the poles and zeros gives no reason to reject the tentative conclusion that $m = 4$.

Table 11. Order Tests for Example 3

order	Mean Sq. Residual		Det(Q)		Normal det	
	Full	Reduced	Full	Reduced	Full	Reduced
1	.482E-5	.560E-5	.262E-3	.312E-3	1.0	1.0
2	.300E-5	.336E-5	.180E-6	.236E-6	.98	.80
3	.293E-5	.333E-5	.233E-9	.131E-9	3.8	2.7
4	.287E-5	.273E-5	.199E-11	.152E-10	25.	30.
5	.280E-5	.260E-5	.208E-16	.134E-13	.3	307.
6	.280E-5	.252E-5	.167E-16	.581E-16	1.7E4	1.1E4

In summary, the normalized $\det(Q_c)$ test is the most effective followed by the $\det(Q_c)$, F-, and mean squared residual tests; the poles and zeros are useful as checks on the other tests. It is best to use all the tests and to use them on several sets of parameter estimates because, consistnet test results imply confidence about both the chosen order and the chosen model type. Also, the computations are very inexpensive: the entire computation time for all tests and parameter estimates for the power example was under 8 seconds on a CDC 6500 computer.

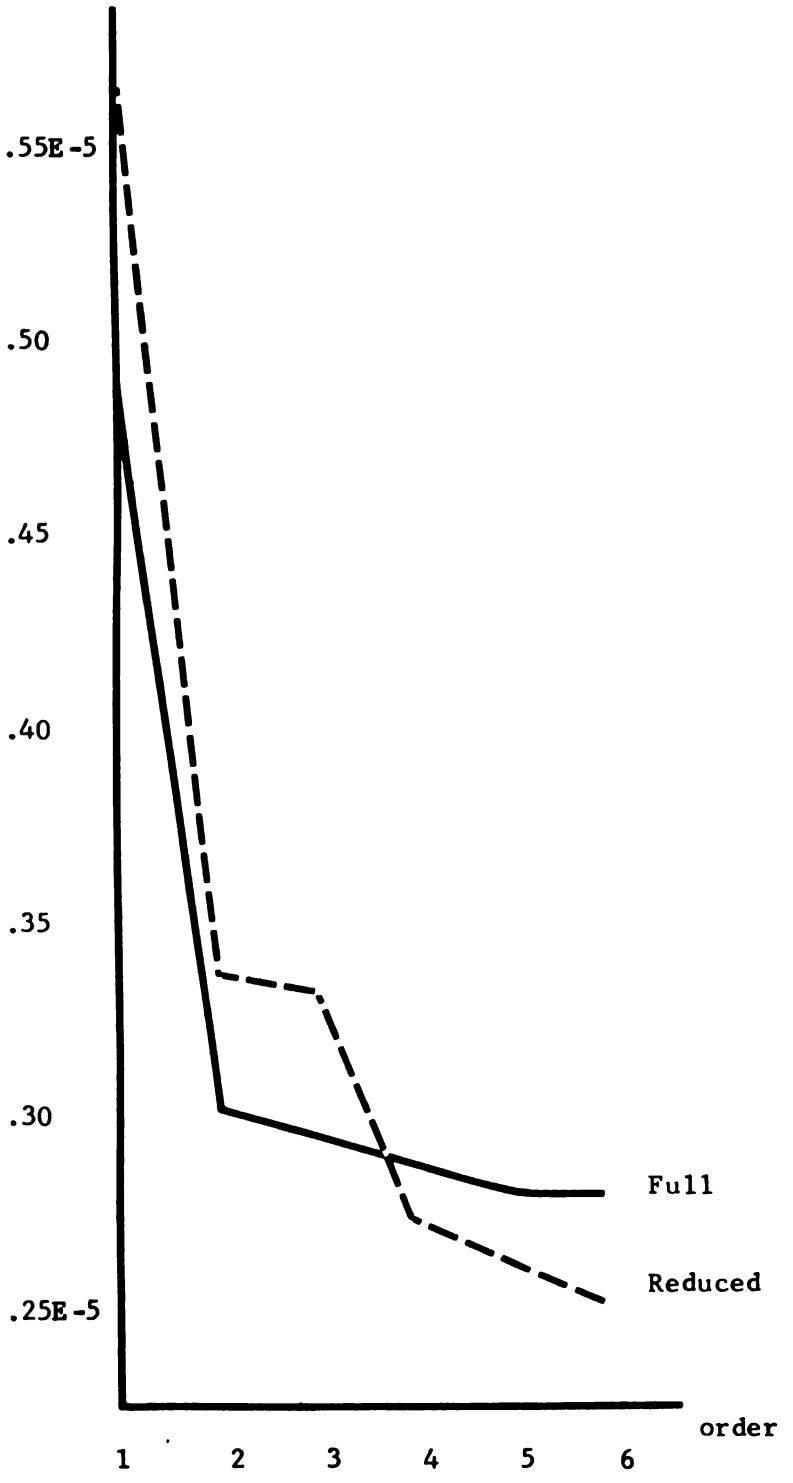

Figure 8. Mean Squared Residual -- Example 3

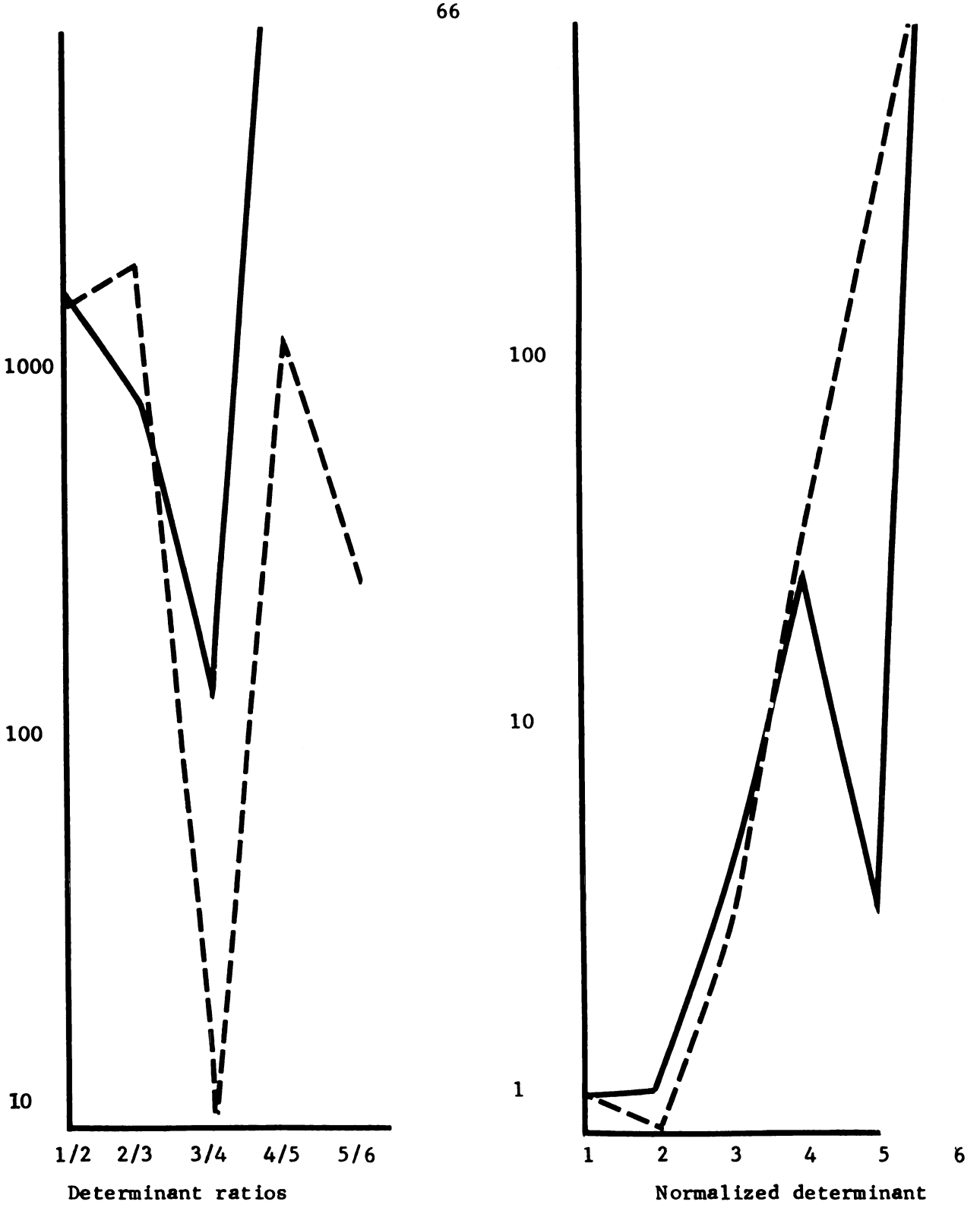

Figure 9. Controllability Determinants -- Example 3

Table 12. F-test for Example 3

Order	1	2	3	4	5
2	116.				
3	61.6	4.65			
4	43.1	4.39	4.06		
5	34.3	4.67	4.60	5.05	
6	27.3	3.52	3.09	2.58	.124
	Full matrix				
2	17.0				
3	8.36	.248			
4	8.21	2.70	5.10		
5	6.46	2.17	3.11	1.10	
6	5.24	1.79	2.29	.899	.708

Table 13. Poles and Zeros for Example 3

Order	Poles -- full matrix				
2	1.000135			-.601	
3	1.000106	.052		-.649	
4	1.000132	.335	(-.450, .037)		
5	1.000110	.416	(-.180, .269)	-.618	
6	1.000110	.514	(-.534, .143)	(-.004, .403)	
order	zeros -- full matrix				Steady state
2			.404		1.858
3	-2.12		.682		1.858
4	-2.63	(.555, .223)			1.858
5	-1.87	(.438, .622)	.780		1.857
6	-1.83	(.442, .623)	.794	-.033	1.857
order	Poles -- reduced matrix				
2	1.000158			-.630	
3	1.000142	-.002		-.613	
4	1.000240	.718	(-.663, .417)		
5	1.000196		(-.747, .462)	(.444, .287)	
6	1.000209	.685	(-.811, .450)	(.157, .548)	
order	zeros -- reduced matrix				Steady state
2			.432		1.810
3	-1.065		.630		1.830
4	-1.992	(.666, .207)			1.827
5	-1.719	(.588, .248)		.185	1.831
6	-1.325	(.497, .433)	.736	-.514	1.834

CHAPTER IV

CONCLUSIONS

4.1 Thesis Results

Choosing a model type is always a difficult and important problem in modeling. Even if it is possible to use the simplest type of dynamic system -- linear, time-invariant, single-input, single-output -- there is still the problem of selecting an order of the model. We have discussed the model order problem for such systems. There are two aspects of this problem: the method of parameter identification and tests to determine the order of the system on the basis of the parameter estimates.

The basic method used in the parameter identification phase was least squares. Least squares was chosen because it requires less knowledge and fewer assumptions than other methods. After introducing the reduced variant of the algorithm, large and small sample properties were derived assuming that the order of the system was known. Some theoretical results when the model order was greater than the system order were given. While reduced least squares often gave better estimates than the full matrix version, seomtimes they were worse. It was definitely useful to have both sets of estimates when estimating the system order.

Several tests to determine the system order were given. Among them are some new ones based on the characteristics of dynamic

systems. The most promising of these is based on the fact that the minimal order of a system equals the rank of the product of the controllability and observability matrices of any representation. Using a representation where the observability matrix is the identity, the test is to look at the determinant of the controllability matrices for successive model orders normalized by the product of the corresponding input coefficients. The normalized determinants are relatively constant when the model order is less than the system order; when they are equal, there is a large jump in the determinant value; and the behavior is erratic when the model order is more than the system order. It was helpful to also use an F-test on the residuals.

4.2 Future Development

Any new area contains many voids between the already known facts. In identification, much work can be done on the theoretical properties when the model order is more than the system order. This will be a very difficult problem because uniqueness is absent. Selection of weighting schemes is also not well developed.

Additional model order tests must be developed. For example, other state-space realizations might generate determinants which have more useful properties than the one we used. Finding useful properties of the determinants is also an area for research. And the general problem of selecting model properties is wide open.

APPENDIX

APPENDIX

A.1 Proof of Theorem 2

The assumptions imply that $C = \lim C_N$ as defined by (2.2.1) exists and is positive definite (see (M1) for details). Intuitively, this makes sense because C is the correlation matrix for the input and output.

$$\text{(A.1.1)} \quad \lim B_N = \lim \begin{matrix} \frac{1}{N}\Sigma_1^N y_{n+i-1}e_{n+i} \\ \vdots \\ \frac{1}{N}\Sigma_1^N y_i e_{n+i} \\ \frac{1}{N}\Sigma_1^N u_{n+i-1}e_{n+i} \\ \vdots \\ \frac{1}{N}\Sigma_1^N u_i e_{n+i} \end{matrix} = \begin{matrix} E(y_{t-i}e_t) \\ \vdots \\ E(y_{t-n}e_t) \\ E(u_{t-1}e_t) \\ \vdots \\ E(u_{t-n}e_t) \end{matrix} = B$$

B exists by assumptions 1) - 3). Let

$$\text{(A.1.2)} \quad Z_N = \frac{1}{N}\phi_N^T\phi_N(\theta_N - \theta) = \frac{1}{N}(\phi_N^T Y_N - (\phi_N^T\phi_N(Y_N - E_N)))$$

$$= \frac{1}{N}\phi_N^T E_N = B_N$$

Using (A.1.1) and (A.1.2) $E(Z_N) = E(B_N) = B$. Also, $E(Z_N Z_N^T) = \frac{1}{N}E(\frac{1}{N}\phi_N^T E_N E_N^T \phi_N)$, a moment matrix which exists. We now look at the limits. Obviously,

$$\text{(A.1.3)} \quad \lim E(C_N\theta_N') = \lim E(Z_N) = B$$

Although $E(C_N \theta_N') \neq E(C_N)E(\theta_N')$, still

$$\text{(A.1.4)} \qquad \lim E(C_N \theta_N') = \lim E(C_N) \times \lim E(\theta_N') = C \lim E(\theta_N')$$

because the stochastic matrices C_N approach the deterministic matrix C. Since C is invertible,

$$\text{(A.1.5)} \qquad \lim E(\theta_N') = C^{-1}B$$

In Theorem 1, B = 0, so the lse is unbiased.

To show mean square convergence, we must show that

$$\text{(A.1.6)} \qquad \lim E(\theta_N'^T \theta_N') = 0$$

Since $\theta_N' = C_N^{-1} Z_N$ and since C_N is symmetric,

$$\text{(A.1.7)} \qquad E(\theta_N'^T \theta_N') = E(Z_N^T C_N^{-2} Z_N)$$

$$\text{(A.1.8)} \qquad \lim E(Z_N^T Z_N) = \lim E(\mathrm{tr}(Z_N Z_N^T))$$

$$= \lim \frac{1}{N} \mathrm{tr}(\text{moment matrix}) = 0$$

Since C is positive definite, C_N^{-2} acts as a weighting matrix for N sufficiently large. Thus,

$$\text{(A.1.9)} \quad \lim E(\theta_N'^T \theta_N') = \lim \frac{1}{N} \mathrm{tr}(\text{weighted moment matrix}) = 0$$

and mean square convergence is proved.

A.2 Proof of Theorem 4

Lep $p = 1$. We will show that $-1 \le -g \le 1$. From (4.2.)

$$-g = \frac{\theta_1^T \theta_o}{\theta_1^T \theta_1} \tag{A.2.1}$$

From (4.2.6) and (4.2.7)

$$\theta_1^T \theta_1 = 1 + \theta_0^T \theta_o \tag{A.2.2}$$

Let $\theta_1^T \theta_o \ge 0$. If $\theta_1^T \theta_o \ge \theta_1^T \theta_1$, then

$$2\theta_1^T \theta_o \ge \theta_1^T \theta_1 + \theta_o^T \theta_o \tag{A.2.3}$$

or

$$0 \ge (\theta_1 - \theta_o)^T (\theta_1 - \theta_o) \tag{A.2.4}$$

which is impossible. The proof is similar for $\theta_1^T \theta_o < 0$. Now assume that for $p = p_o$ we have

$$G_0 = -(H_o H_o^T)^{-1} H_o \theta_o = -L_o^{-1} J_o \tag{A.2.5}$$

We will express G_1, corresponding to $p = p_o + 1$, in terms of G_o. Notice that $(L_o)_{i,j} = \theta_i^T \theta_j = f_{|i-j|}$, allowing us to write

$$L_1 = \begin{bmatrix} L_o & K \\ K^T & f_o \end{bmatrix} \tag{A.2.6}$$

where

$$K = \mathrm{col}(f_{p-1}, f_{p-2}, \ldots, f_1) \tag{A.2.7}$$

and

$$J_1 = \mathrm{col}(J_o, f_p) \tag{A.2.8}$$

L_1^{-1} can be expressed in terms of L_o^{-1} as (F1)

$$(A.2.9) \qquad L_1^{-1} = \begin{bmatrix} I & -L_o^{-1}K \\ 0 & 1 \end{bmatrix} \begin{bmatrix} L_o^{-1} & 0 \\ 0 & (f_o - K^T L_o^{-1} K)^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -K^T L_o^{-1} & 1 \end{bmatrix}$$

Finally, we get

$$(A.2.10) \qquad G_1 = \begin{bmatrix} G_o - dL_0^{-1}K \\ d \end{bmatrix} = \begin{bmatrix} g_1^o + dg_{p-1}^o \\ g_2^o + dg_{p-2}^o \\ \vdots \\ g_{p-1}^o + dg_1^o \\ d \end{bmatrix}$$

where g_i^o is the i-th element of G_o and

$$(A.2.11) \qquad d = -(f_o - K^T L_o^{-1} K)^{-1} (f_p - K^T L_o^{-1} G_o)$$

To show that the roots of (2.4.) are either all within or all outside the unit circle, we use the Lehmer-Schur method (R1). Define

$$(A.2.12) \qquad P_o(z) = z^{p-1} + g_1^o z^{p-2} + \ldots + g_{p-1}^o$$

and

$$(A.2.13) \qquad P_o^*(z) = g_{p-1}^o z^{p-1} + \ldots + 1$$

The roots of P^* are the inverse conjugates of the roots of P, i.e. if all the roots of P are inside the unit circle then all the roots of P^* are outside. In a similar manner $P_1(z)$ and $P_1^*(z)$ are defined from G_1. We now form

$$(A.2.14) \quad T(P_1(z)) = dP_1(z) - P_1^*(z) = (d^2-1)g_{p-1}^o z^{p-1} + \ldots + (d^2-1)$$
$$= (d^2-1)P_o^*(z)$$

The Lehmer-Schur method states that if $T(P_1(0)) < 0$, then $T(P_1(z)) = \text{constant} \cdot P_0^*(z)$ and $P_1^*(z)$ have the same number of zeros within the unit circle; if $T(P_1(0)) > 0$, then $P_o^*(z)$ and $P_1(z)$ have the same number of zeros inside. In summary, if $|d| \leq 1$ and G_o defines a stable system, then G_1 will also define a stable system; if $|d| > 1$ and G_o is stable, then every mode of G_1 will be unstable; etc., and the theorem is proved.

It would be more interesting if we could prove that $|d| < 1$. It is not too difficult to show that $d > -1$, but the other inequality is more more difficult.

A.3 Steady-state Response to a Unit-step Input

We will show that the steady-state response of the system (1.1.1) to a unit-step input is

(A.3.1) $$B/(1 - A)$$

where

(A.3.2) $$A = \Sigma_1^n a_i \qquad B = \Sigma_1^n b_i$$

We will show this in both the time domain (difference equation representation) and z-domain (transfer function representation) to illustrate the differences of the two domains.

In the time domain, $u_t = 1$ for all $t = 0,1,\ldots$ Then

$$\begin{aligned} y_1 &= b_1 \\ y_2 &= a_1 y_1 + (b_1 + b_2) \\ &\vdots \\ y_{n+1} &= a_1 y_n + \ldots + a_n y_1 + B \\ &\vdots \\ y_{n+k} &= \Sigma_1^n a_i y_{n+k-i} + B \end{aligned}$$

Taking limits,

(A.3.3) $$y_{ss} = \Sigma_1^n a_i y_{ss} + B = y_{ss} A + B$$

and the result is shown.

In the z-domain, the transfer function is

(A.3.4) $$T(z) = Y(z)/U(z)$$

and the unit-step response is (01)

(A.3.5) $$Y(z) = G(z)\,\frac{z}{z-1}$$

The final value theorem states

(A.3.6) $$y_{ss} = \lim_{t\to\infty} y_t = \lim_{z\to 1} ((z-1)Y(z)).$$

Thus,

(A.3.7) $$y_{ss} = \lim_{z\to 1} (zG(z)) = \frac{B}{1-A}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

A1 Anderson, J.H., "Geometrical Approach to Reduction of Dynamical Systems", Proc. IEEE, Vol. 114, July 7, 1967, pp. 1014-1018.

A2 Astron, K.J., Lectures on the Identification Problem -- The Least Squares Method, Lund Inst. of Tech., Div. of Automatic Control Report 6906, 1968.

A3 Astrom, K.J. and Eykoff, P., "System Identification -- A Survey", Automatica, Vol. 7, pp. 123-162.

A4 Astrom, K.J., Bohlin, T. and Wensmark, S., Automatic Construction of Linear Stochastic Dynamic Models for Stationary Industrial Processes with Random Disturbances Using Operating Records, IBM Nordic Laboratory Technical Paper TP 18.150, 1965.

A5 Athans, M. and Falb, P.L., Optimal Control, McGraw-Hill, NY, 1965.

B1 Bacopoulos, A. and Gaff, B., "On the Reduction of a Problem of Minimization of n+1 Variables to a Problem of One Variable", SIAM J. on Numer. Anal., Vol. 8, No. 1, pp. 97-103.

B2 Bellman, R. and Astrom, K.J., "On Structural Identifiability", Mathematical Biosciences, Vol. 7, 1969, pp. 329-339.

B3 Bevington, P.R., Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, NY, 1969.

B4 Boling, R., personal communication.

B5 Brockett, R.W., Finite Dimensional Linear Systems, Wiley, NY, 1970.

C1 Central Electricity Generating Board, "Demand Response to Changes of Voltage", Report PL-ST/1/70, London, 1970.

C2 Chen, C.T., "A New Look at Transfer Function Design", Proc. IEEE, Vol. 59, No. 11, November 1971, pp. 1580-1585.

C3 Cheney, E.W., "Numerical Problems of Linear Algebra", Mimeographed lecture notes, Michigan State University, 1969.

D1 Davison, E.J., "A Method for Simplifying Linear Dynamic Systems", IEEE Auto. Cont., Vol. 11, No. 1, January 1966, pp. 93-101.

D2 Draper, N.R. and Smith, H., Applied Regression Analysis, Wiley, NY, 1966.

E1 Efroymson, M.A., "Multiple Regression Analysis, in Mathematical Methods for Digital Computers, A. Ralston and H.S. Wilf, eds., Wiley, NY, 1960.

E2 Elgerd, O.I., Electric Energy Systems: An Introduction, McGraw-Hill, NY, 1971.

F1 Farrar, D.E. and Glauber, R.R., "Multicollinearity in Regression Analysis: the Problem Revisited", Rev. Econ. and Stat., Vol. 49, February 1967.

F2 Forsythe, G.E., "Generation and Use of Orthogonal Polynomials for Data-fitting on a Digital Computer", J. SIAM, Vol. 5, 1957, pp. 74-88.

F3 Frame, J.S., "Matrix Operations and Generalized Inverses", IEEE Spectrum, March 1964.

G1 Gilbert, E.G., "Controllability and Observability in Multivariable Control Systems", SIAM J. on Control, Vol. 2, No. 1, 1963, pp. 128-151.

G2 Goodman, D. and Hiller, J., "Determination of System Order", First Hawaii International Conference on Systems Science, 1968, pp. 527-530.

G3 Gopinath, B., "On the Identification and Control of Linear Systems", Ph.D. thesis, Stanford University, 1968.

H1 Haitovsky, Y., "Multicollinearity in Regression Analysis: Comment", Rev. Econ. and Stat., Vol. 51, No. 4, 1969, pp. 486-490.

H2 Hurwicz, L., "Least Squares Bias in Time Series", in Statistical Inference in Dynamic Economic Systems, T.C. Koopmans, ed., Cowles Commission monograph No. 10, Wiley, NY, 1950.

I1 IBM, "System/360 Scientific Subroutine Package Programmer's Manual", 1966.

I2 Isaacson, E. and Keller, H.B., Analysis of Numerical Methods, Wiley, NY, 1966.

J1 Jenkins, G.M. and Watts, D.G., Spectral Analysis and its Applications, Holden-Day, San Francisco, 1968.

K1 Kalman, R.E., "Mathematical Description of Linear Dynamical Systems", SIAM J. on Control, Vol. 2, No. 1, 1963, pp. 152-192.

K2 Kalman, R.E., Falb, P.L. and Arbib, M.A., Topics in Mathematical Systems Theory, McGraw-Hill, NY, 1970.

L1 Lewis, T.O. and Odell, P.L., Estimation in Linear Models, Prentice-Hall, Englewood Cliffs, NJ, 1971.

M1 Mann, H.B. and Wald, A., "On the Statistical Treatment of Linear Stochastic Difference Equations", Econometrica, Vol. 11, 1943, pp. 173-220.

M2 Marshall, S.A., "An Approximate Method for Reducing the Order of a Linear System", Control, December 1966, pp. 642-643.

M3 Monahan, D.D., System Identification by Stochastic Approximation, Ph.D. thesis, Michigan State University, 1970.

O1 Ogata, K., Modern Control Engineering, Prentice-Hall, NY, 1970.

P1 Park, G.L., Monahan, D.D. and Schwartz, M.W., "An Application of a Stochastic-Approximation, System-Identification Algorithm to Interconnected Electric Power Systems", Midwest Power Symposium, Ames, Iowa, 1970.

P2 Powell, M.J.D., Numerical Methods for Unconstrained Optimization", SIAM Review, Vol. 12, No. 1, January 1970.

R1 Ralston, A., A First Course in Numerical Analysis, McGraw-Hill, NY, 1965.

R2 Rosenbrock, H.H., State Space and Multivariable Theory, Wiley, NY, 1970.

S1 Soderstrom, T., Notes on Pseudoinverses: Applications to Identification, Lund Inst. of Tech. Div. of Auto Control, Report 7003, 1970.

S2 Staley, R.M. and Yue, P.C., "On System Parameter Identifiability", Information Sciences, Vol. 2, pp. 127-138.

T1 Theil, H., Economic Forecasts and Policy, North-Holland, Amsterdam, 1961.

T2 Tintner, G., Econometrics, Wiley, NY, 1952

W1 Werther, M.H.F., Recursive Algorithm for the Best Approximate Solution of Linear Equations, Management Information Services, Detroit, 1970.

W2 Wold, H. and Jureen, L., Demand Analysis, Wiley, NY, 1953.

W3 Woodside, C.M., "Estimation of the Order of Linear Systems", Automatica, Vol. 7, No. 6, 1971.

Z1 Zadeh, L.A. and Desoer, C.A., Linear System Theory: The State Space Approach, McGraw-Hill, NY, 1963.

Z2 Zadeh, L.A. and Polak, E., eds., System Theory, McGraw-Hill, NY, 1969.